文章编号：0253-2239(2009)04-1117-05

基于近红外的 Fisher 分类法识别茶叶原料品种的研究

周 健[1] 成 浩[1] 叶 阳[1] 王丽鸳[1] 贺 巍[2] 刘 栩[1] 陆文渊[1]

([1]中国农业科学院茶叶研究所茶树资源与改良研究中心，国家茶树改良中心，浙江 杭州 310008；[2]南京农业大学，江苏 南京 210095)

摘要 提出一种可对成品茶的原料品种进行准确识别的方法。在实验中对不同原料品种(龙井 43# 与其他品种)制成的茶叶样本进行近红外光谱的采集，通过主成分分析(principal component analysis，PCA)后获得了 20 个主成分，利用逐步回归法筛选出 8 个主成分作为自变量，建立茶叶原料品种的 Fisher 识别函数对成品茶的原料品种进行识别分析。实验结果表明建立的识别函数能很好地对茶叶的原料品种进行准确识别，在定标集中的识别准确率达到了 96.8%，并且利用外部样本进行验证的识别准确率也达到了 93.5%。本实验证实了利用 PCA 和 Fisher 识别组合分析识别成品茶原料品种的可行性。

关键词 光学鉴别；茶；原料品种识别；近红外；逐步回归

中图分类号 O657.33 **文献标识码** A **doi**：10.3788/AOS20092904.1117

Recognition for Raw Material Cultivar of Manufactured Tea With Fisher Discriminant Classification With Principal Components Analysis

Zhou Jian[1] Cheng Hao[1] Ye Yang[1] Wang Liyuan[1] He Wei[2]
Liu Xu[1] Lu Wenyuan[1]

([1] *Research Center for Tea Germplasm and Improvement, National Center for Tea Improvement, Tea Research Institute, Chinese Academy of Agriculture Science, Hangzhou, Zhejiang* 310008, *China*
[2] *Nanjing Agricultural University; Nanjing, Jiangsun* 210095, *China*)

Abstract A new method for recognition of raw material cultivar of manufactured tea with Fisher discriminant classification using near infrared spectra. In this study, spectra of samples with different raw material cultivar (Longjing43# and other cultivar) were collected. 20 principal components were obtained by PCA. 8 Principal components by step wise was used to establish the Fisher function for discriminant classification to recognize the raw material cultivar of manufactured tea. The result showed that the function performed well in recognition of raw material cultivar of manufactured tea. The accuracy for recognition was 96.8% in calibration set. 93.5% was obtained for unknown samples in test set. The result proved that it was feasible to recognize the raw material cultivar with combined analysis of PCA and Fisher discriminant classification using near infrared spectra.

Key words optical discrimination; tea; recognition for raw material cultivar; near infrared; stepwise regression

1 引 言

由于茶叶的原料品种的不同，加工成的成品茶的品质差异很大。某些特定品种制成的茶叶由于其优异的品质而受到消费者的欢迎，在市场上的销售状况往往好于普通品种的茶叶，如由白茶品种制成的安吉白茶，由大红袍品种制成的乌龙茶——大红袍，龙井 43# 品种制成的西湖龙井等。但是由于利益的驱使这些特定品种制成的茶叶受到假冒产品的困扰，特别是一些

收稿日期：2008-04-22；**收到修改稿日期**：2008-09-27

基金项目：国家科技基础条件平台项目(2005DKA1002)和浙江省自然科学基金(Y305096)资助课题。

作者简介：周 健(1979－)，男，助理研究员，硕士，主要从事茶树生物技术方面的研究工作。E-mail：zjph263@126.com

栽培面积小、产量少、品质优异的乌龙茶名种如大红袍等[1]。由于加工工艺的相同,特殊品种与普通品种制成的茶叶的区别十分困难。因此,利用对原料品种的识别来实现成品茶的真伪鉴定是一个可能的途径。

近红外光谱分析技术是一种快速、无损、绿色的分析方法,已经广泛应用于石油化工、农业、食品和饮料、药物定量定性分析等领域[2~7]。许多研究证明利用合适的分析方法结合近红外可识别成品茶的种类,但都是对于不同加工工艺茶叶的区分,鉴别相对容易[8~14]。Fisher 线性识别作为一种分类方法已经被成功的广泛使用[15~18],但是还未有利用 Fisher 与 PCA 组合分析成品茶的近红外光谱进行成品茶的原料品种识别的报道。

本实验以龙井 43# 与其他的品种的原料制成的成品茶为实验样本,通过对样本的近红外光谱进行 PCA 分析,采用逐步回归的方法提取符合要求的主成分作为自变量,建立 Fisher 分类识别函数,分析函数对于龙井 43# 和其他样本制成的成品茶分类判别的准确性,从而建立起一种成品茶的原料品种的准确识别方法,为实现特定品种原料制成的茶叶的品种溯源奠定基础。

2 实验部分

2.1 实验材料

本实验采用的样品为龙井 43# 品种的鲜叶原料与其他品种的鲜叶原料(迎霜、乌牛早、鸠坑种等 10 个品种)制成的成品茶叶(均采用龙井茶的加工工艺)。其中龙井 43# 为原料品种的样本为 33 份(组 1),其他品种的样本为 61 份(组 2),共 94 份。将样本随机排列,每 3 个样本中选择 2 个作为定标样本,剩下的 1 个作为验证样本验证模型对未知样本的识别效果,最终 94 个样本分为两部分:63 个样本作为定标集(包括 21 个龙井 43# 样本和 42 个其他品种样本)和 31 个样本为验证集(包括 12 个龙井 43# 样本和 19 个其他品种样本),所有样品直接现场取自杭州市西湖龙井产区的茶农及厂家,其他地区的扁形茶样品也是直接到产地的生产厂家直接获取,保证了样品来源品种的真实和可靠(表 1)。

2.2 样品处理

在进行近红外光谱的采集前,所有的样品均经过粉碎处理,具体过程为:称取大约 20 g 左右的茶样,放入中药粉碎机(DFT-50,20000 r/min,浙江林大机械厂生产)粉碎约 30 s 左右,之后将磨碎后的粉末过 40 目筛,然后准确称取 10 g 作为近红外的分析材料。

表 1 定标集与验证集样本数量

Table 1 Sample amount of raw material cultivar in calibration set and test set

Raw material cultivar	Sample amount in calibration set /piece	Sample amount in test set /piece	Total /piece
Longjing43 /piece	21	12	33
Other cultivar/piece	42	19	61
Total /piece	63	31	94

2.3 近红外光谱的采集

样品的近红外漫反射光谱的采集在 IFS 28/N (Bruker,德国)近红外光谱仪上进行,积分球漫反射检测器,扫描次数 64,分辨率 3.857 cm^{-1},光谱采集软件软件为 Opus Quant 2,扫描区域为 10000~3500 cm^{-1}。数据点的间隔为 3.857 cm^{-1},因此采集的光谱的数据点为 1946 个。采集时室温控制在 25 ℃左右,湿度保持稳定。

2.4 主成分分析与识别函数的建立

对所有样本进行 PCA 分析,提取前 20 个主成分,采用逐步回归的方法构建 Fisher 判别函数,每次选入 Wilk's λ 统计量最小的变量进入函数进行逐步回归运算,根据变量进入函数的偏 F 检验值最小值为 3.84,剔除函数的最大值为 2.71 的标准(SPSS 回归分析时的软件默认标准,3.84 为卡方检验的临界值,F 值小于 3.84 则在 0.05 水平没有显著差异),即当被加入的变量的值 $F \geqslant 3.84$ 时,该变量进入函数,当要被移出的变量的值 $F \leqslant 2.71$ 时,该变量被移出函数,最终建立两个识别函数 F_1 与 F_2。

2.5 函数对于品种的识别效果分析

利用 2.4 的方法可获得两个识别函数 F_1 与 F_2,根据 Fisher 分类的判别标准。若 $F_1 > F_2$,则归入组 1,若 $F_1 < F_2$,则为组 2。按照此标准分别分析识别函数对于定标集,交叉验证和未知样本的原料品种的识别准确率,确定函数对成品茶原料品种识别效果。

2.6 数据分析

利用近红外光谱仪的随机软件 Opus quant2 采集光谱数据,采用软件 Unscrambler95 (CAMO, OLSO, 挪威)上进行 PCA 分析,SPSS 10.0(SPSS,美国)进行 Fisher 识别函数的建立和识别效果的分析。

3 结果与分析

3.1 PCA 分析

图 1 为所有样本的近红外光谱,图中表明所有的近红外光谱在 7200~4000 cm^{-1} 的区域有丰富的

吸收峰，说明这个区域含有丰富的信息量；在 11000～7200 cm^{-1}区域信息很弱，主要反映的是一些分子结构相对简单的分子结构信息，4000 cm^{-1}以下的光谱主要贡献噪声，因此选取信息量丰富的 4000～7200 cm^{-1}区段的光谱作为分析对象。对所有样本的近红外光谱进行 PCA 分析获得所有样本的前 20 个主成分，结果表明前 20 个主成分能解释 99.9% 以上的光谱信息(图 2)。考虑到不同原料品种的制成的成品茶的近红外光谱差异都很小，若选取主成分数过少的话可能会造成识别的误差，为了获得较高的识别准确率并且减少计算量，选取前 20 个主成分作为建立识别函数的分析对象。

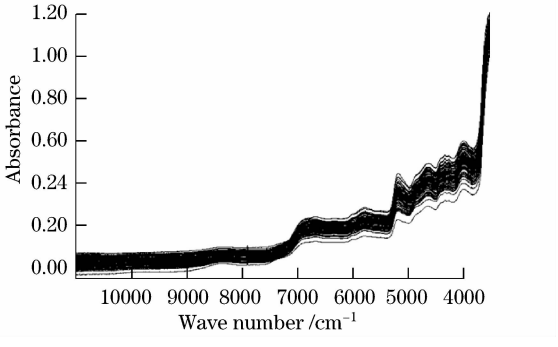

图 1　样本的近红外光谱

Fig. 1　Near infrared absorbance spectra of all samples

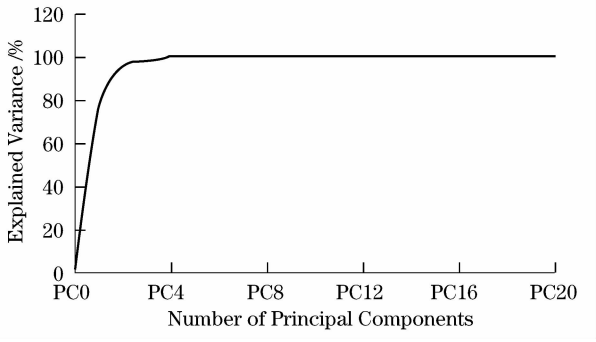

图 2　主成分分析获得的前 20 个主成分对光谱信息的解释程度

ig. 2　Accumulative reliabilities plot of the first 20 principal components calculated from the spectra of PCA

3.2　Fisher 自变量的确立

在建立识别函数的过程中，利用逐步回归的运算方法分别选择变量进入方程，每次回归运算时 Wilks λ 统计量最小的变量被选入方程，最终确定 8 个主成分的 F 值的符合变量选入函数的 F 检验值最小值为 3.84 的标准，可作为判别建立判别函数的变量(表 2)。而其他主成分则不满足 $F \geqslant 3.84$ 的选入函数的标准，不作为建立判别函数的变量。

3.3　原料品种的识别函数的建立

利用 SPSS 中的分类判别分析功能建立 Fisher 识别函数 F_1 和 F_2，若将样本的相应主成分分别代入函数 F_1 和 F_2，若计算结果为 $F_1 > F_2$，则为组 1(龙井 43 品种样本)；若 $F_1 < F_2$，则为组 2(其他品种样本)。

表 2　构成函数的 8 个主成分在选入函数时的 Wilks λ 和 F 检验值

Table 2　Wilks λ and F value of the 8 principal components when they were selected to establish Fisher's discriminant function

	PC_7	PC_{13}	PC_{11}	PC_4	PC_{18}	PC_{14}	PC_3	PC_1
Order for entering the discriminant function	1	2	3	4	5	6	7	8
Wilk λ statistic	0.582	0.506	0.446	0.392	0.361	0.331	0.309	0.287
F Statistic	86.657	11.759	11.224	11.927	5.843	4.816	4.086	4.019

$$F_1 = -3.195 + 0.955PC_1 + 5.950PC_3 + 25.387PC_4 - 477.984PC_7 - 547.828PC_{11} - 874.253PC_{13} - 566.430PC_{14} + 1285.103PC_{18}$$

$$F_2 = -1.203 - 0.512PC_1 - 2.629PC_3 - 19.014PC_4 + 303.538PC_7 + 320.147PC_{11} + 493.298PC_{13} + 393.888PC_{14} - 520.462PC_{18}$$

3.4　识别函数对于定标集的识别效果和交叉验证

利用识别函数 F_1 和 F_2 分别对定标集和样本进行识别，确定识别函数对于定标集样本的分类效果。结果表明，21 个龙井 43# 样本中有 20 个准确识别，识别的准确率为 95.2%，而 42 个其他品种样本中有 40 个识别正确，识别的准确率为 97.6%，总的识别正确率为 96.8%，说明了利用函数对于定标集样本的原料品种分类具有良好的识别效果(表 3)。

表 3　Fisher's 识别函数对于定标集的分类效果

Table 3　Classification of samples in calibration set with established Fisher's discriminant function

	Actual Group	Predicted Group		Total
	Result	1	2	
Amount	1	20	1	21
	2	1	41	42
%	1	95.2	4.8	100.0
	2	2.4	97.6	100.0
Correctly classified /%			96.8	

3.5　识别函数对于未知样本的原料品种的识别效果

为确定识别函数对于外部未知样本的原料品种的识别效果，将 31 个未知样本的经过 PCA 分析的相应主成分代入识别函数 F_1 与 F_2，比较 F_1 和 F_2

的计算结果，根据 $F_1 > F_2$，归入组 1，若 $F_1 < F_2$，归入组 2 的分类标准，将分类结果与实际结果进行比较，确定识别函数对未知茶叶样本的原料品种的识别效果。结果表明，31 个样本中 2 个样本识别错误，一个龙井 43# 品种的样本被错误分入其他品种的样本中，一个其他品种样本被错误识别为龙井 43# 样本，其余 29 个样本均被识别正确，识别的正确率为 93.5%，说明识别函数也同样适用于对外部茶叶样本的原料品种的识别，证实了利用 Fisher 识别函数进行成品茶原料品种识别的可行性(表 4)。

表 4 Fisher's 识别函数对于验证集未知样本的分类

Table 4 Classification of samples in test set with established Fisher's discriminant function

	Original set	F_1	F_2	Predicted set		Original set	F_1	F_2	Predicted set
1	1	8.030	2.097	1	17	2	−5.139	2.141	2
2	1	−0.182	−1.686	1	18	2	−8.227	−2.849	2
3	1	0.936	−1.059	1	19	2	−8.505	−0.720	2
4	1	0.580	−0.416	1	20	2	−2.295	−0.113	2
5	1	−0.371	0.456	2*	21	2	−3.417	−0.617	2
6	1	2.939	1.747	1	22	2	−1.574	2.202	2
7	1	3.908	−1.105	1	23	2	−5.275	−2.264	2
8	1	−0.461	−3.784	1	24	2	−9.099	−3.887	2
9	1	4.381	0.262	1	25	2	−3.931	−1.152	2
10	1	4.116	−0.081	1	26	2	−7.279	−1.604	2
11	1	1.926	0.384	1	27	2	2.914	−0.966	1*
12	1	2.761	−0.134	1	28	2	−4.240	−1.943	2
13	2	−6.092	1.000	2	29	2	−5.042	−0.203	2
14	2	−7.072	0.052	2	30	2	−6.475	−4.207	2
15	2	−4.722	−1.916	2	31	2	−4.712	−1.202	2
16	2	−2.948	−0.129	2					

*：错误识别样本 incorrectly classified samples

4 讨 论

以不同原料品种制成的成品茶的近红外光谱进行 PCA 分析获得的前 20 个主成分作为分析对象，利用逐步回归的方法建立了针对龙井 43# 的品种制成的成品茶的原料品种的 Fisher 识别函数，实现了对成品茶原料品种的准确识别，在定标集和外部样本验证集中的识别准确率都达到了 96.8% 和 93.5%。因此本实验证实了利用 PCA 和 Fisher 组合分析识别特定品种制成的成品茶的可行性，对于实现优异和珍稀品种制成的名优茶的识别与保护都有重要的积极意义。

参 考 文 献

1 Ye Naixing, Zheng Naihui. Original Producing area protection of Fujian famous tea varieties[J]. *J. Fujian Agriculture and Forestry and University* (Nature science edition), 2004, **33**(4): 459～462
叶乃兴，郑乃辉. 福建名茶与原产地保护[J]. 福建农林大学学报，2004，**33**(4)：459～462

2 Zou Xiaobo, Zhao Jiewen. Methods of characteristic wavelength region and wavelength selection based on genetic algorithm[J]. *Acta Optica Sinica*, 2007, **27**(7): 1316～1321
邹小波，赵杰文. 用遗传算法快速提取近红外光谱特征区域与特征波长[J]. 光学学报，2007，**27**(7)：1316～1321

3 Liu Yande, Chen Xingmiao, Ouyang Aiguo. Non-destructive measurement of soluble solid content in gannan navel oranges by visible/near-infrared spectroscopy[J]. *Acta Optica Sinica*, 2008, **28**(3): 478～481
刘燕德，陈兴苗，欧阳爱国. 黄酒糖度预测的可见近红外光谱方法研究[J]. 光学学报，2008，**28**(3)：478～481

4 Liu Fei, He Yong, Wang Li. Methods for the prediction of sugar content of rice wine using visible-near infrared spectroscopy[J]. *Acta Optica Sinica*, 2007, **27**(11): 2054～2058
刘 飞，何 勇，王 莉. 黄酒糖度预测的可见近红外光谱方法研究[J]. 光学学报，2007，**27**(11)：2054～2058

5 T Davies. The history of near infrared spectroscopic analysis: Past, present and future. From sleeping technique to the morning star of spectroscopy[J]. *Analysis*, 1998, **26**(4): 17～19

6 M. Blanco, I. Villarroya. NIR spectroscopy: a rapid-response analytical tool[J]. *Trends in Analytical Chemistry*, 2002, **21**(4): 240～250

7 Candolfi A, De Maesschalck R, Jouan-Rimbaud D *et al*.. The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra [J]. *J. Pharmaceutical & Biomedical Analysis*, 1999, (21): 115～132

8 Chen Quansheng, Zhao Jiewen, Zhang Haidong *et al*.. Identification of authenticity of tea with near infrared spectroscopy based on support vector machine[J]. *Acta Optica Sinica*, 2006, **26**(6): 933～937
陈全胜，赵杰文，张海东 等. 基于支持向量机的近红外光谱鉴别茶叶的真伪[J]. 光学学报，2006，**26**(6)：933～937

9 Quansheng Chen, Jiewen Zhao, C. H. Fang *et al*.. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM)[J]. *Spectrochimica Acta Part A*, 2007, **66**: 568～574

10 Li Xiaoli, He Yong, Qiu Zhengjun. Application of PCA-ANN method to fast discrimination of tea varieties using near infrared Spectroscopy[J]. *Spectroscopy and Spectral Analysis*, 2007, **27**(2): 279～282
李晓丽，何　勇，裘正军. 一种基于可见近红外光谱快速鉴别茶叶品种的新方法[J]. 光谱学与光谱分析，2007，**27**(2)：279～282

11 Chen Quansheng, Zhao Jiewen, Zhang Haidong *et al*.. Application of near infrared reflectance spectroscopy to the identification of tea using SIMCA pattern recognition method[J]. *J. Food Science*, 2006, **27**(4): 186～189
陈全胜，赵杰文，张海东 等. SIMCA 模式识别方法在近红外光谱识别茶叶中的应用[J]. 食品科学，2006，**27**(4)：186～189

12 Jiewen Zhao, Quansheng Chen, Xingyi Huang *et al*.. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine[J]. *J. Pharmaceutical and Biomedical Analysis*, 2006, **41**: 1198～1204

13 J. Luypaert, M H. Zhang, D L. Massart. Feasibility study for the use of near infrared spectroscopy in thequalitative and quantitative analysis of green tea[J]. *Camellia sinensis* (*L.*) *Analytica Chimica Acta*, 2003, **478**: 303～312

14 Yong He, Xiaoli Li, Xunfei Deng. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model[J]. *J. Food Engineering*, 2007, **79**: 1238～1242

15 Y. S. Qu, B. L. Adam, M. Thornquist *et al*.. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data[J]. *Biometrics*, 2003, **59**(1)143～151

16 S. A Billings, K L Lee. Nonlinear fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm[J]. *Neural Netw*, 2002, **15**(2): 263～270

17 Gavin C. Cawley, Nicola L. C. Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers [J]. *Pattern Recognition*, 2003, **36**(11): 2585～2592

18 T Van Gestel, J. A. K. Suykens, G lanckriet *et al*.. Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel Fisher discriminant analysis[J]. *Neural Comput*, 2002, **14**(5): 1115～1147