

一种基于可见-近红外光谱快速鉴别茶叶品种的新方法

李晓丽, 何 勇, 裘正军*

浙江大学生物工程与食品科学学院, 浙江 杭州 310029

摘 要 提出了一种用可见-近红外光谱技术快速无损鉴别茶叶品种的新方法。应用可见-近红外光谱仪测定 5 个品种茶叶的光谱曲线, 用主成分分析法对不同品种茶叶进行聚类分析并获得茶叶的可见-近红外光谱数据的主成分, 再结合人工神经网络技术建立模型进行品种鉴别。主成分分析表明, 以主成分 1 和 2 对所有建模样本的得分值做出的得分图, 对不同种类茶叶具有较好的聚类作用, 可以定性分析茶叶种类。把主成分分析得到的前 6 个主成分作为神经网络的输入, 茶叶品种值作为神经网络的输出, 通过 5 个茶叶品种共 125 个样本的训练和学习, 建立了茶叶品种鉴别的 3 层 BP 人工神经网络模型, 对未知的 25 个样本进行鉴别, 品种识别准确率达到 100 %。说明本文提出的方法具有很好的分类和鉴别作用, 为茶叶的品种快速鉴别提供了一种新方法。

关键词 可见-近红外光谱; 茶叶; 品种; 主成分分析; 人工神经网络; 鉴别

中图分类号: S123; TH744.1 **文献标识码:** A **文章编号:** 1000-0593(2007)02-0279-04

引 言

茶叶于 2000 多年前发源于中国, 目前已经传播到全世界。茶是我国的传统饮料, 也是世界三大饮料之一。茶叶主要分为 3 种, 即不发酵的绿茶、半发酵的乌龙茶以及发酵的黑茶和红茶。绿茶是我国茶叶重点出口茶类, 出口量占我国出口茶叶总量的 70 %, 占世界绿茶总量的 85 % 以上。茶叶作为健康饮料风靡世界。随着国际贸易的发展和人们饮食结构的调整, 对茶叶质量的要求越来越高, 全面提高茶叶产品的质量, 已日趋为人们所重视。茶叶质量安全市场准入制度即将实施, 我国茶叶将在国际大市场中运营, 这些都对我国茶叶加工业提出更高的科技创新要求。用简便、快速、准确的现代分析方法定量鉴别其品种和等级成为分析工作者的一项重要任务。本文以光谱技术为基础研究茶叶的品种鉴别。

现代近红外光谱分析技术, 可充分利用全谱段或多波长下的光谱数据进行定性或定量分析。由于近红外光谱分析技术具有速度快、效率高、成本低、测试重现性好、测量方便等特点, 已经被越来越多地应用于食品工业、石油化工、制药工业等领域。国内外很多学者研究利用近红外光谱技术区别物质品种, 如咖啡品种^[1]、草莓品种^[2]、瓜类品种^[3]、大豆品种^[4]、道地山药^[5]等, 但还没有人应用可见-近红外光谱

对茶叶品种进行鉴别^[6]。

BP 神经网络模型是一个强有力的学习系统, 能够实现输入与输出之间的高度非线性映射。目前使用最多的是多层结构的误差反向传播学习算法(BP), 并且已经证明此种模型可以逼近任何连续的非线性曲线^[7]。主成分分析是多元统计中的一种数据挖掘技术。在不丢失主要光谱信息的前提下, 选择为数较少的新变量来代替原来较多的变量, 解决了由于谱带的重叠而无法分析的困难。我们将主成分分析(PCA)和基于误差反向传播算法(back propagation, BP)多层前馈神经网络有机地结合起来建立了 5 个品种茶叶的近红外光谱鉴别模型。

1 材料与方法

1.1 仪器设备

实验使用美国 ASD (analytical spectral device) 公司的 Handheld Field Spec 光谱仪, 其光谱采样间隔(波段宽) 1.5 nm, 测定范围 325 ~ 1 075 nm, 扫描次数 30 次, 分辨率 3.5 nm, 探头视场角为 25 °。光源采用与光谱仪配套的 14.5 V 卤素灯。光谱数据分析软件为 ASD View Spec Pro, Unscramble V9.1 和 DPS (data procession system for practical statistics)。

收稿日期: 2005-12-12, 修订日期: 2006-03-26

基金项目: 国家自然科学基金项目(30671213, 30600371), 高等学校博士学科点专项科研基金课题(20040335034), 高等学校优秀青年教师教学科研奖励计划(02411)和浙江省重大科技攻关项目(2005C12029)资助

作者简介: 李晓丽, 女, 1982 年生, 浙江大学生物系统工程与食品科学学院博士研究生

e-mail: yhe @zju.edu.cn

*通讯联系人

1.2 样品来源及光谱的获取

买来市售 5 种茶叶分别是西湖龙井、浙江龙井、羊岩勾青、雪水云绿、庐山云雾, 各取 30 个样本, 共计 150 个样本。茶叶均用直径为 65 mm 高度为 14 mm 培养皿盛装。装满一个培养皿茶叶作为一个实验样本。全部样本随机分成建模集和预测集, 建模集有 125 个样本, 预测集有 25 个样本。光谱仪经白板校准后进行测试。光谱仪置于茶叶样本的上方, 距茶叶表面 120 mm, 探头视场角为 25°, 对每一个样本扫描 30 次。

Table 1 Varieties of the sample in the research

品种	类别	产地	样本数目/个
西湖龙井	绿茶	杭州龙井	30
浙江龙井	绿茶	浙江	30
羊岩勾青	绿茶	浙江临海	30
雪水云绿	绿茶	浙江桐庐	30
庐山云雾	绿茶	江西庐山	30

1.3 光谱数据预处理

为了去除来自高频随机噪音、基线漂移、样本不均匀、

光散射等影响, 需要进行光谱预处理。采用平均平滑法, 选用平滑窗口大小为 9, 此时能很好滤除各种因素产生的高频噪音, 再进行 MSC(multiplicative scatter correction)处理。由于光谱曲线在首端和末端有较大噪音, 如图 1 所示, 所以只取 400 ~ 1 000 nm 波段的光谱用于分析^[8,9]。

1.4 人工神经网络模型

建立了一个 3 层的人工神经网络结构, 各层传递函数都用 S 型(Sigmoid)函数。网络输入层节点数为 6, 经多次实验确定隐含层节点数为 8, 输出层节点数为 1。

2 试验结果与分析

2.1 茶叶样本的近红外漫反射光谱

5 个品种茶叶的典型近红外光谱曲线如图 1 所示。图 1 中横坐标为波长, 范围是 325 ~ 1 075 nm, 纵坐标为光谱漫反射率。从图 1 中可以看出, 不同品种茶叶的光谱曲线有明显区别, 并具有一定的特征性和指纹性, 这一差异为茶叶的不同品种鉴别奠定了数学基础。选择波长范围在 400 ~ 1 000 nm 的光谱, 应用 ASD View Spec Pro 软件, 并转换成 ASCII 码, 形成反射率矩阵, 用主成分分析法对其聚类。

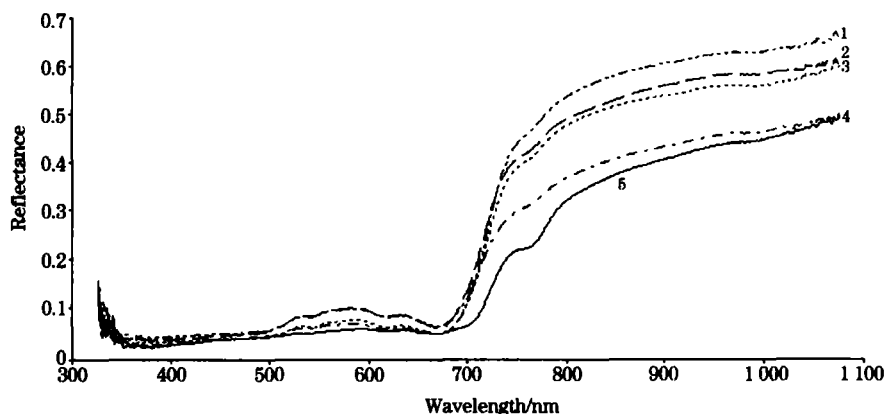


Fig 1 Near infrared reflectance spectra of three different types of tea

1: 浙江龙井; 2: 西湖龙井; 3: 雪水云绿; 4: 羊岩勾青; 5: 庐山云雾

2.2 主成分分析对不同品种茶叶进行定性分析

主成分分析的目的是将数据降维, 以消除众多信息共存中相互重叠的信息部分。通过对原始大量光谱变量进行转换, 使数目较少的新变量成为原变量的线性组合, 而且, 新变量能最大限度的表征原变量的数据结构特征, 并不丢失信息。对 5 种茶叶共 150 个样本进行主成分分析聚类。分析表明前 2 个主成分的累计贡献率达到 96.378%, 对 5 种茶叶有一定的聚类作用, 能定性区别不同品种茶叶。

图 2 表示 150 个样本的主成分 1 和 2 得分图, 图中横坐标表示每个样本的第一主成分得分值, 纵坐标表示每个样本的第二主成分得分值。图 2 中羊岩勾青、雪水云绿、庐山云雾、龙井明显分成 4 类, 但是龙井中的西湖龙井和浙江龙井有部分重叠在一起。图 2 说明主成分 1 和 2 对茶叶有很好的聚类作用, 尤其是对羊岩勾青、雪水云绿、庐山云雾、龙井 4 种茶叶有很好的聚类作用。从图 2 中可以看出, 庐山云雾的 30 个样本聚合度较好, 紧密地分布在图 2 的第一象限, 其他

样本大都处于第一象限之外。雪水云绿的聚合度也很好, 所有 30 个样本大都位于第三象限。羊岩勾青茶叶大部分样本位于第三象限, 但是有一些样本落在了第一象限。西湖龙井和浙江龙井紧密聚合在一起, 大部分分布在第二象限, 但是有部分样本落在了第三象限。西湖龙井与浙江龙井的分界线不如前面几种清晰, 分析原因在于它们属于同一茶叶品种, 前 2 个主成分还不能很好地区分。我们可以看出 5 种茶叶在图 2 中呈条状聚类, 5 种茶叶的样本点分布在 5 个平行的聚类块中, 当主成分 1 增大时, 主成分 2 减小; 当主成分 1 减小时, 主成分 2 增大。即它们的主成分 1 和 2 负相关。于是我们可以发现主成分 1 和 2 把 5 个品种茶叶聚类成 5 个相互平行的条状块, 它们之间的分界线(除了西湖龙井和浙江龙井)很明显, 但是对于西湖龙井和浙江龙井的鉴别作用还不显著。所以光定性分析是不够的, 下面通过建立定量模型来鉴别茶叶品种。

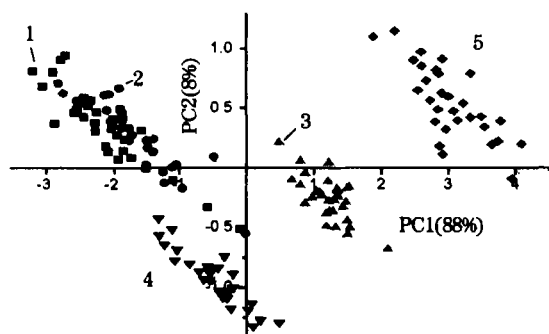


Fig 2 PCA scores plots(PC1 ×PC2) for tea sample across the entire spectral region

1: 西湖龙井; 2: 浙江龙井; 3: 羊岩勾青;
4: 雪水云绿; 5: 庐山云雾

2.3 基于神经网络定量分析建立茶叶品种鉴别模型

光谱波段从 400 ~ 1 000 nm 共有 600 个点, 但是, 采用 600 个点计算时, 计算量大, 而且有些区域样品的光谱信息很弱, 与样品的组成或性质间缺乏相关关系。所以通过主成分分析, 选取对于茶叶品种敏感的新变量作为输入建立神经网络品种鉴别模型。可大大简化模型, 提高计算速度和分类准确度。图 3 表示前 10 个主成分对原始变量的解释程度。图 3 中横坐标表示主成分数目, 顺序从左到右依次为 PC1 到 PC10。纵坐标表示主成分的累积可信度, 即主成分能够解释原始波长变量的程度。如图 3 可知, 前 6 个主成分能够解释原始波长变量的 99.941%, 说明前 6 个主成分可代表原来 600 个变量, 说明主成分分析是一种非常有效的数据挖掘方法, 它把原来的 600 个波长变量压缩成了彼此正交的新变量, 这 6 个新变量彼此间是互不影响的, 而且能代表绝大部分原变量包含的信息。

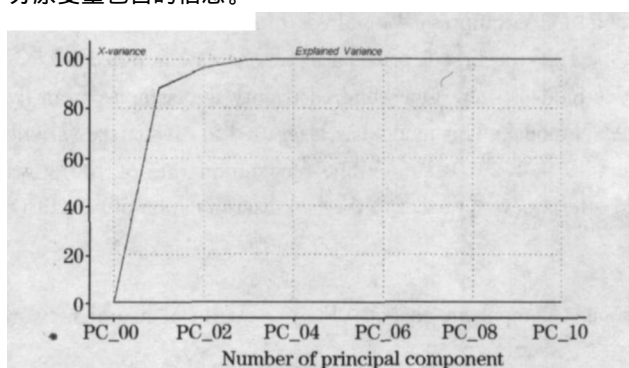


Fig 3 Accumulative reliabilities plot of the first 10 principal components calculated from the plant spectra of PCA model

把前 6 主成分作为 BP 神经网络的输入变量建立鉴别模型, 用符号 1, 2, 3, 4, 5 来分别代表西湖龙井、浙江龙井、羊岩勾青、雪水云绿、庐山云雾等 5 种茶叶的品种值, 通过调整隐含层的节点数来优化网络结构^[10], 经过反复试验得到最佳网络结构为 6 个输入单元, 8 个隐含单元和 1 个输出单元的 3 层 BP 神经网络模型, 分别对 125 个建模样本进行

学习, 经 3000 次的训练迭代后, 拟合残差为 1.267×10^{-4} , 对未知的 25 个样本进行预测, 预测结果如表 2 所示。模型对 5 个品种茶叶的判断均正确, 预测准确率为 100%。尤其是西湖龙井和浙江龙井, 在前面的主成分聚类中很难区分, 但是 BP 神经网络能够准确的鉴别 2 种茶叶, 如表 2 中前 10 号样本所示。之前的一些品种鉴别只给出了定性结果, 没有建立模型定量鉴别。即使某些研究中建立了预测模型, 模型的正确识别率也不高, 如孙素琴等对道地山药进行鉴别的正确识别率为 70%^[15]。

Table 2 Prediction result for 25 unknown samples by BP-ANN model

预测样本序号	真实值	预测值
(1)	1	1.005 05
(2)	1	1.004 04
(3)	1	1.003 39
(4)	1	1.004 78
(5)	1	1.003 49
(6)	2	2.119 56
(7)	2	1.674 71
(8)	2	1.846 01
(9)	2	2.217 80
(10)	2	1.569 95
(11)	3	3.119 60
(12)	3	2.996 52
(13)	3	3.013 43
(14)	3	2.913 51
(15)	3	2.991 98
(16)	4	4.099 82
(17)	4	3.910 86
(18)	4	3.597 91
(19)	4	3.760 66
(20)	4	4.028 05
(21)	5	4.994 60
(22)	5	4.978 31
(23)	5	4.990 44
(24)	5	4.992 53
(25)	5	4.984 15

Note: (1)-(5), 西湖龙井; (6)-(10), 浙江龙井; (11)-(15), 羊岩勾青; (16)-(20), 雪水云绿; (21)-(25), 庐山云雾。

3 结 论

通过实验获得了 5 个品种茶叶的光谱特征曲线, 利用这些光谱信息结合主成分分析人工神经网络建立了茶叶品种鉴别的模型, 该模型的预测效果很好, 识别率达到 100%。说明运用可见-近红外光谱技术可以快速、准确、无损的对茶叶品种进行鉴别。我们用于茶叶品种分析的可见-近红外光谱在 325 ~ 1 075 nm 范围内, 说明该波长范围是对茶叶品种敏感的特征波段。本文提出的主成分结合人工神经网络方法, 用于品种鉴别得到的识别率得到很大提高, 说明该模型用于鉴别茶叶品种是成功的, 为茶叶品种的快速无损检测提供了一种新的方法。

参 考 文 献

- [1] Esteban-Diez I, Gonzalez-Saiz J M, Pizarro C. *Analytica Chimica Acta*, 2004, 514: 57.
- [2] Lopez M. *Near Infrared Spectroscopy*, Proceedings of the 10th International Conference, Chichester, UK: NIR Publications, 2002. 335.
- [3] Zsolt Seregely, Tamas Deak, Gyorgy Denes Bisztray. *Chemometrics and Intelligent Laboratory Systems*, 2004, 72: 195.
- [4] Turza S, Toth A, Varadi M. *Near Infrared Spectroscopy: Proceedings of the 8th International Conference*, Chichester, UK: NIR Publications, 1998. 183.
- [5] SUN Su-qin, TANG Jun-ming, YUAN Zi-min, et al(孙素琴, 汤俊明, 袁子民, 等). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2003, 23(2): 258.
- [6] LUO Yi-fan, GUO Zhen-fei, ZHU Zhen-yu, et al(罗一帆, 郭振飞, 朱振宇, 等). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2005, 25(8): 1230.
- [7] HE Yong, LI Xiao-li, SHAO Yong-ni(何 勇, 李晓丽, 邵咏妮). *Lecture Notes in Computer Science*, 2005, 3809: 1053.
- [8] QI Xiao-ming, ZHANG Lu-da, DU Xiao-lin, et al(齐小明, 张录达, 杜晓林, 等). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2003, 23(5): 870.
- [9] YIN Qiu, SHU Xiao-zhou, XU Zhao-an, et al(尹 球, 疏小舟, 徐兆安, 等). *Journal of Infrared and Millimeter Waves(红外与毫米波学报)*, 2004, 23(6): 427.
- [10] ZHAO Chen, QU Hai-bin, CHENG Yi-yu(赵 琛, 瞿海斌, 程翼宇). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2004, 24(1): 50.

Application PCA-ANN Method to Fast Discrimination of Tea Varieties Using Visible/ Near Infrared Spectroscopy

LI Xiao-li, HE Yong, QIU Zheng-jun *

College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China

Abstract A new method for the discrimination of varieties of tea by means of visible/ near infrared spectroscopy (Vis/ NIRS) (325-1 075 nm) was developed. A relation has been established between the reflectance spectra and the tea varieties. The data set consists of a total of 150 samples of tea. First, the data was analyzed with principal component analysis(PCA). It appeared to provide the reasonable clustering of the varieties of tea. Meanwhile PCA compressed hundreds of spectral data into a small quantity of principal components which described the body of the spectra; the first 6 principal components computed by PCA were applied as inputs to a back propagation neural network with one hidden layer. One hundred twenty five samples from five varieties were selected randomly, then they were used to build BP-ANN model. This model has been used to predict the varieties of 25 unknown samples; the residual error for the calibration samples is 1.267×10^{-4} . The recognition rate of 100 % was achieved. This model is reliable and practicable. So this paper could offer a new approach to the fast discrimination of varieties of tea.

Keywords Visible-near infrared spectroscopy; Tea; Variety; Principal component analysis (PCA); Artificial neural network (ANN); Discrimination

(Received Dec. 12, 2005; accepted Mar. 26, 2006)

* Corresponding author