文章编号： 1672-8785(2012)03-0044-05

近红外光谱和模式识别技术在西湖龙井与浙江龙井茶叶鉴别中的应用

张 龙[1,2] 王飞娟[2] 潘家荣[2] 朱 诚[2,1]

(1. 浙江大学生命科学学院植物生理学与生物化学国家重点实验室，浙江 杭州 310058；
2. 中国计量学院生命科学学院，浙江 杭州 310018)

摘 要：为了鉴别西湖龙井和浙江龙井茶叶，采用近红外光谱分析技术结合化学计量学方法建立了识别模型。先对原始光谱进行标准正态变换 (Standard Normal Variant, SNV) 预处理，然后分别采用最小二乘判别分析 (Partial Least Square Regression-discriminant Analysis, PLS–DA) 、最小二乘支持向量机 (Least Square Support Vector Machine, LSSVM) 和径向基人工神经网络 (Radial Basis Function Neural Network, RBFNN) 三种模型对西湖龙井和浙江龙井茶叶进行预测。最小二乘支持向量机参数通过网格搜索和完全交叉验证得到优化。经优化后，惩罚系数 (γ) 和核函数参数 (δ^2) 分别为 229.1 和 124.9；RBFNN 最佳隐藏层神经元个数为 27 个。通过比较可知，LSSVM 的预测性能最好，其校正集均方根误差 (RMSECV) 和相关系数 (R^2) 分别为 0 和 1，验证集均方根误差 (RMSEP) 和相关系数 (R^2) 也分别为 0 和 1，分辨正确率为 100 %。

关键词：浙江龙井茶叶；西湖龙井茶叶；近红外光谱；偏最小二乘回归判别分析；最小二乘支持向量机；径向基神经网络

中图分类号： TS201.2 **文献标识码**： A **DOI**: 10.3969/j.issn.1672-8785.2012.03.009

Application of Near Infrared Spectroscopy and Pattern Recognition Model in Discrimination of Xihu Longjing Tea and Zhejiang Longjing Tea

ZHANG Long[1,2], WANG Fei-juan[2], PAN Jia-rong[2], ZHU Cheng[2,1]

(1. State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou 310058, China;
2. College of Life Sciences, China Jiliang University, Hangzhou 310018, China)

Abstract: To discriminate Xihu Longjing Tea and Zhejiang Longjing Tea, a recognition model was established by using a near infrared spectroscopy combined with chemometrics. First, the raw spectra were preprocessed with the Standard Normal Variant (SNV) function. Then, the Xihu Longjing Tea and Zhejiang Longjing Tea were predicted respectively with three models of Partial Least Square Regression-

收稿日期：2012–02–12
基金项目：浙江省重点科技创新团队——农产品安全标准与检测技术创新团队项目 (2010R50028)；“十一五”国家科技支撑计划项目：食品安全关键技术——粮油、蔬果等安全控制技术的研究 (Y3100246)
作者简介：张龙 (1986-)，男，山东临沂人，博士研究生，主要研究方向为农产品产地溯源。E-mail: 10907017@zju.edu.cn

discriminant analysis (PLS-DA), Least Square Support Vector Machine (LSSVM) and Radial Basis Function Neural Network (RBFNN). The parameters of the LSSVM were optimized via grid search and leave-one-out cross-validation technologies. After optimization, the penalty coefficient (γ) and kernel function (δ^2) were 229.1 and 124.9 respectively. The best number of the neuron in hidden layer of RBFNN was 27. By comparison, LSSVM had the best prediction performance. Its RMSECV and R^2 in the calibration set were 0 and 1 respectively while its RMSEP and R^2 in the validation set were also 0 and 1 respectively. As a result, the Xihu Longjing Tea and Zhejiang Longjing Tea were classified correctly.

Key words: Zhejiang longjing tea; Xihu longjing tea; near infrared spectroscopy; partial least square regression-discriminant analysis; least square support vector machine; radial basis function neural network

0 引言

龙井茶在我国具有 1200 多年的历史，是我国最著名的名优茶。国家质量监督检验检疫总局在 2001 年第 28 号公告中规定了龙井茶的原产地域产品保护范围。将龙井茶原产地域划分为西湖、钱塘和越州三个产区。其中，用西湖产区的茶鲜叶生产的龙井茶被称为西湖龙井茶；浙江龙井茶是指除杭州市西湖区之外的浙江省其他市县生产的龙井茶。西湖龙井和浙江龙井茶在茶叶品种和加工工艺上往往相同，导致它们在外形、香气和滋味等方面都很难加以区分。因此，人们需要一种可以快速鉴定西湖龙井茶的方法来保护西湖龙井茶的品牌，规范西湖龙井和浙江龙井茶叶的销售市场，从而形成良性的竞争环境。

近红外分析技术是一种具有快速高效、操作简便、成本低和重现性好等特点的无损检测技术，它在农产品、食品、药品以及工业领域都有广泛应用 [1]。目前，近红外分析技术主要用于判别农产品与食品的掺假情况 (鉴别蜂蜜 [3]、小麦 [4] 和螺旋藻 [5] 等)，实现对橄榄油 [6] 的产地溯源以及预测牛奶的蛋白质含量 [7]、谷物和玉米的水分 [8−9] 及蛋白 [10−11] 含量等定量分析。近红外光谱吸收带是有机物质中 C–H、O–H 和 N–H 键基频吸收的倍频、合频和差频以及吸收的叠加 [2]。不同基团产生的光谱在吸收峰的位置和强度上有所不同。根据朗伯 – 比尔定律，随着样品组成或者结构的改变，其光谱特征也将发生变化，即光谱吸收强度与组分含量之间存在一定的数学关系。这种数学关系很复杂，所以一般通过数学模型来建立，以便于实现对样品的定性或定量分析。

本文通过分析浙江龙井和西湖龙井茶叶的近红外光谱指纹特征，并借助最小二乘判别分析 (Partial Least Square Regression-discriminant Analysis, PLS–DA)、偏最小二乘支持向量机 (Least Square Support Vector Machine, LSSVM) 以及径向基人工神经网络 (Radial Basis Function Neural Network, RBFNN) 等化学计量学分析方法，建立浙江龙井和西湖龙井茶叶的判别模型。我们旨在寻找出一种可以快速有效鉴别浙江龙井和西湖龙井茶叶的方法。

1 材料和方法

1.1 样品采集与处理

本试验采用的茶叶样品为杭州地区梅家坞、龙井村、杨梅岭、上满觉陇、下满觉陇、翁家山、四眼井、葛衙营、慈母桥和梅外里等地的西湖龙井以及浙江省富阳、新昌、诸暨、嵊州和绍兴县等不同产地的浙江龙井茶共 60 份 (详细信息见表 1)。其中，西湖龙井茶叶样品为 30 份，浙江龙井茶样品为 30 份。

取 10 g 茶叶样品放入粉碎机中粉碎 10 min，然后将其过 180 目筛后用于近红外光谱扫描。

1.2 光谱采集

将研磨过筛后的样品置于近红外光谱仪器的专用石英开口样品杯中，并将样品摊匀压实，然后开始采集其漫反射光谱图。设置扫描次数为 32 次，分辨率为 2 cm^{-1}，光谱波数范围为 12000 ∼ 4000 cm^{-1}。采集时，室温控制在 25 ℃左右，湿度保持稳定。对每个样品进行重复装样，扫描 3 次；取 3 次的平均值作为光谱数据值。

1.3 数学模型

在西湖龙井和浙江龙井茶叶样品组中每组随机选取 20 个样品作为校正集，每组剩余 10 个

表 1 样品采集信息表

西湖龙井茶叶	样品数	西湖龙井茶叶	样品数	浙江龙井茶叶	样品数
梅家坞村	6	翁家山村	3	富阳	9
龙井村	6	四眼井村	1	诸暨	5
杨梅岭村	2	梅外里	3	新昌	8
上满觉陇村	1	葛衙庄	5	嵊州	6
下满觉陇村	2	慈母桥	1	绍兴	2

样品作为验证集。

1.3.1 偏最小二乘判别分析法

偏最小二乘判别分析法 (PLS–DA) 是一种基于偏最小二乘算法的判别分析方法。偏最小二乘法是由瑞典统计学家沃尔德提出的 [12]，它可以用于建立一个或多个因变量与自变量之间的回归关系。 PLS 法最先被用于建立连续响应变量模型, 即通过指定每一个样品的归类构建响应变量 (哑变量)，然后建立响应变量 (哑变量) 与解释变量 (光谱变量) 的回归预测模型，最后通过比较响应变量的预测值大小来确定样本的归类 [13]。

1.3.2 最小二乘支持向量机

支持向量机 (Support Vector Machine, SVM) 的学习方法是结构风险最小化，其优化问题的约束条件是训练误差，优化目标是置信范围值最小化。最小二乘支持向量机是基于支持向量机方法的一种改进算法。它采用最小二乘线性系统作为损失函数，而不是像传统的支持向量机那样采用二次规划方法。其求解速度较快，在各个领域中都得到了广泛的应用和进一步的研究发展。

1.3.3 径向基人工神经网络

径向基人工神经网络是一种两层前向型神经网络，它包含一个具有径向基函数神经元的隐层和一个具有线性神经元的输出层，其网络结构见图 1 。

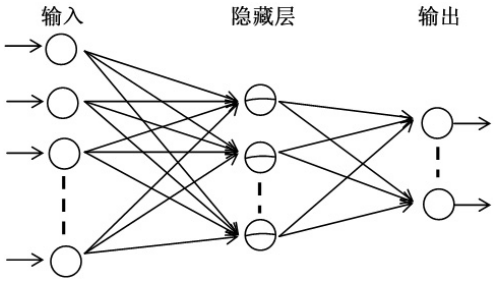


图 1 径向基神经网络的结构图

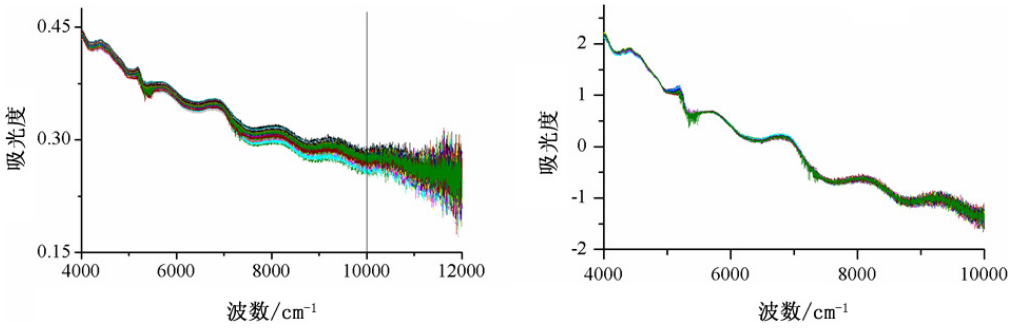
1.4 数据分析

本文在光谱预处理、模型参数选择以及模型性能评价中均采用 Matlab V7.0 进行分析。

2 结果和讨论

2.1 光谱预处理

西湖龙井和浙江龙井茶叶在 10000 ~ 12000 cm^{-1} 近红外光谱波段上的噪声信息繁多。因此，



(a) 原始谱图 (b) 标准正态变换预处理后得到的谱图

图 2 龙井茶叶的近红外光谱图

在下一步的分析中只选用 4000 ~ 10000 cm^{-1} 光谱波段 (见图 2)。近红外光谱降噪方法有多种多样。在实验中要根据样本状态是液体还是固体来选择合适的光谱预处理方法。由于茶叶粉末的光谱噪声主要是由粒径的大小差异等因素引起的，本文采用 SNV 光谱预处理方法。

主成分分析是常用的数据降维和特征提取方法，它可以提高计算的简便性和增强预测的稳定性。图 3 为主成分解释变量累积图。其中，前 10 个主成分只反映了 58.2 % 的原始光谱信息，但是其判别正确率达到了 100 %。在偏最小二乘回归判别分析中选用的主成分数为 10 个。本文将主成分提取的前 10 个主成分的得分作为 LSSVM 和 RBFNN 的输入。

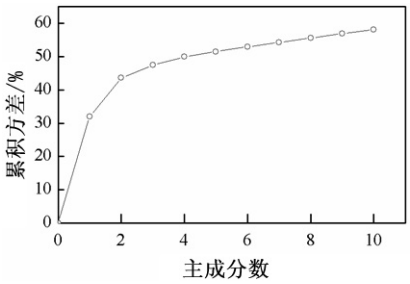


图 3　主成分解释变量累积图

2.2　模型参数选择

为西湖龙井和浙江龙井茶叶建立三种模型，并将西湖龙井茶叶赋值为 2，浙江龙井茶叶赋值为 1。以近红外光谱变量 (PLS–DA) 或主成分得分 (RBFNN 和 LSSVM) 为 X 变量，赋值为分类变量建立模型，分别研究其对西湖龙井和浙江龙井茶叶的预测能力。当预测值大于 1.5 时，将此样本识别为西湖龙井茶叶；当预测值小于 1.5 时，将此样本识别为浙江龙井茶叶。

2.2.1　最小二乘支持向量机

本文中，最小二乘支持向量机的核函数是径向基函数。核函数参数 δ^2 需要优化；此模型中的惩罚系数 (γ) 也需要优化。将完全交叉验证作为参数寻优，参数寻优则选择网格搜索技术进行 (见图 4)。结果表明，惩罚系数 (γ) 和核函数参数 (δ^2) 分别为 229.1 和 124.9。

2.2.2　径向基人工神经网络

径向基人工神经网络隐藏层的作用函数也是径向基函数。它具有很强的非线性估计能力。本文以训练集交叉验证误差均方根 (RMSECV) 为指标对径向基人工神经网络隐藏层的神经元数进行选择。当隐藏层的最佳神经元个数增加到 27 个时，模型的预测误差下降到 1×10^{-4}，如图 5 所示。

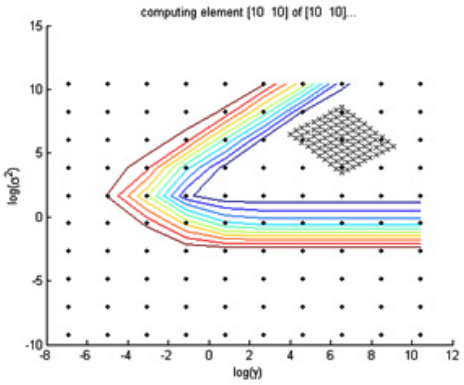


图 4　最小二乘支持向量机的参数寻优步骤

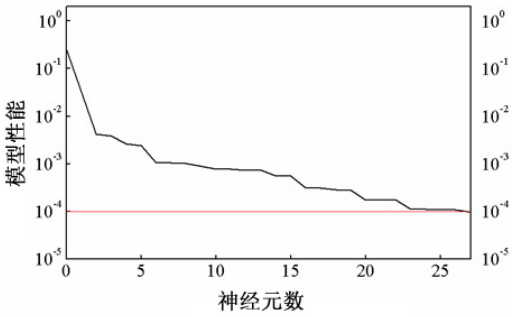


图 5　隐藏层的神经元数对径向基神经网络性能的影响

2.3　模型判别结果的比较

模型性能评价标准采用校正集参数 RMSECV 与 R^2 和验证集参数 RMSEP 与 R^2。从表 2 中可以看出，三种模型都可以实现对西湖龙井和浙江龙井茶叶的 100 % 鉴别，但是 LSSVM 模型的性能最好。其校正集 RMSECV 和 R^2 分别为 0 和 1，验证集 RMSEP 和 R^2 也分别为 0 和 1。

近年来，红外光谱在龙井茶叶识别中的应用已有报道。占茉莉等人采用主成分分析和聚类分析对西湖龙井和浙江地区其他品牌的龙井茶叶进行了研究，基本实现了西湖龙井与浙江龙井茶叶的分类 [14]。周健等人利用偏最小二乘回归实现了对西湖龙井茶叶的 100 % 预测，但是其模型的稳定性仍需进一步优化 [15]。本文比较

表 2 偏最小二乘判别分析法、径向基人工神经网络和最小二乘支持向量机模型的性能比较

	校正集			验证集		
	RMSECV	R^2	CR (%)	RMSEP	R^2	CR (%)
PLS-DA	5.15E-05	1.0000	100.0	0.0409	0.9933	100.0
RBFNN	0.0098	0.9996	100.0	0.1211	0.9414	100.0
LSSVM	0	1.0000	100.0	0	1.0000	100.0

了偏最小二乘判别分析、径向基人工神经网络和最小二乘支持向量机三种模型对西湖龙井和浙江龙井茶叶的预测能力。结果表明，径向基人工神经网络对西湖龙井和浙江龙井茶叶的预测能力最差，而最小二乘支持向量机的预测能力高于线性模型偏最小二乘判别分析。

3 结论

本文将近红外光谱与化学计量学方法相结合，实现了对西湖龙井与浙江龙井茶叶的鉴别。由于近红外光谱数据量大，首先对近红外光谱数据进行降维，然后将降维后的光谱特征信息分别输入偏最小二乘判别分析、径向基人工神经网络和最小二乘支持向量机模型，最后比较不同模型对西湖龙井与浙江龙井茶叶的鉴别能力。结果表明，三种模型对西湖龙井与浙江龙井茶叶的判别率都达到了 100 %。通过比较三种模型的性能参数 RMSECV 、 R_c^2 、 RMSEP 和 R_p^2 可知，最小二乘支持向量机对西湖龙井与浙江龙井茶叶的鉴别能力最强。

参考文献

[1] Williams P, Norris K. Near Infrared Technology in the Agricultural and Food Industries (2nd ed.) [M]. St Paul: American Association of Cereal Chemist, 2001.

[2] 陆婉珍，袁洪福，徐广通，等. **现代近红外光谱分析技术（第一版）** [M]. 北京：中国石化出版社，2000.

[3] Zhu X R, Li S F, Shan Y, et al. Detection of Adulterants Such As Sweeteners Materials in Honey Using Near-infrared Spectroscopy and Chemometrics [J]. *Journal of Food Engineering*, 2010, **101**: 92–97.

[4] Cocchi M, Durante C, Foca G, et al. Durum Wheat Adulteration Detection by NIR Spectroscopy Multivariate Calibration [J]. *Talanta*, 2006, **68**(5): 1505–1511.

[5] Wu D, Nie P C, Cuello J, et al. Application of Visible and Near Infrared Spectroscopy for Rapid and Non-invasive Quantification of Common Adulterants in Spirulina Powder [J]. *Journal of Food Engineering*, 2011, **102**: 278–286.

[6] Lin P, Chen Y M, He Y. Identification of Geographical Origin of Olive Oil Using Visible and Near-infrared Spectroscopy Technique Combined with Chemometrics [J]. *Food Bioprocess Technol*, 2012, **5**: 235–242.

[7] Wu D, He Y, Feng S J, et al. Study on Infrared Spectroscopy Technique for Fast Measurement of Protein Content in Milk Powder Based on LS-SVM [J]. *Journal of Food Engineering*, 2008, **84**(1): 124–131.

[8] 朱之光，熊宁，余敦年，等. GBT 24896-2010. 粮油检验，稻谷水分含量测定 [S]. 北京：中国标准出版社， 2010.

[9] 吴存荣，唐怀建，冯锡仲，等. GBT 24900-2010. 粮油检验，玉米水分含量测定 [S]. 北京：中国标准出版社， 2010.

[10] 唐瑞明，陈洁，吴存荣，等. GBT 24901-2010. 粮油检验，玉米粗蛋白质含量测定 [S]. 北京：中国标准出版社， 2010.

[11] 熊宁，余敦年，刘利，等. GBT 24897-2010. 粮油检验，稻谷粗蛋白质含量测定 [S]. 北京：中国标准出版社， 2010.

[12] Wold S. Pattern Recognition by Means of Disjoint Principle Components Models [J]. *Pattern Recognition*, 1976, **8**: 127–139.

[13] 孙淑敏，郭波莉，魏益民，等. 近红外光谱指纹分析在羊肉产地溯源中的应用 [J]. **光谱学与光谱分析**, 2011, **33**(4): 937–941.

[14] 占茉莉，李勇，魏益民，等. 应用 FT-IR 光谱指纹分析和模式识别技术溯源的研究 [J]. **核农学报**, 2008, **22**(6): 829–833.

[15] 周健，成浩，贺巍，等. 基于近红外的 PLS 量化模型鉴定西湖龙井真伪的研究 [J]. **光谱学与光谱分析**, 2009, **29**(5): 1251–1254.