文章编号：1672-8785(2015)12-0041-06

基于近红外光谱的西湖龙井茶产地的精细判别

张　龙[1]　潘家荣[2]　朱　诚[2]
(1. 丽水学院，浙江 丽水 323000；
2. 中国计量学院，浙江 杭州 310018)

摘　要：不同产区的西湖龙井茶的品质具有差异。采用近红外光谱技术和光谱预处理、主成分分析和判别模型等数学方法鉴别了分别产自龙井村、梅家坞村和葛衙庄三个地区的西湖龙井茶。结果表明，二阶导数光谱预处理方法对去除近红外光谱中的噪音最有效，贝叶斯判别分析是这三个地区产的西湖龙井茶的最佳判别模型。在模型中输入5个主成分数后，最佳的原始判别率和交叉验证判别率分别为100%和82.35%。在交叉验证判别中，产自葛衙庄、龙井村和梅家坞的茶叶的判别正确率分别为80%、83.33%和83.33%。因此，该模型可以用于龙井村、梅家坞村和葛衙庄三个地区产的西湖龙井茶的鉴别，为西湖龙井茶产区的判别提供理论依据。

关键词：西湖龙井；近红外光谱；光谱预处理；主成分分析；判别分析

中图分类号：TN219　　**文献标志码**：A　　**DOI**：10.3969/j.issn.1672-8785.2015.12.008

Precise Discrimination of Xihu Longjing Tea from Different Producing Regions Based on Near-infrared Spectra

ZHANG Long[1], PAN Jia-rong[2], ZHU Cheng[2]
(1. Lishui University, Lishui 323000, China;
2. China Jiliang University, Hangzhou 310018, China)

Abstract: Xihu longjing tea from different producing regions has different quality. Near-infrared spectroscopy, spectral pretreatment, principal component analysis and discriminant model are used to discriminate Xihu longjing tea from Longjing village, Meijiawu village and Geya village. The results show that the second derivative pretreatment method is most effective for the removal of the noise in near infrared spectra and the Bayes discriminant analysis is the best discriminant model for the tea from the above three regions. Setting the components as 5 in the Bayes model, the best original discriminant rate and the cross-validation discriminant rate are 100% and 82.35% respectively. In the cross-validation, the discriminant accuracies of the tea from Longjing village, Meijiawu village and Geya village are 80%, 83.33% and 83.33% respectively. Therefore, the model can be used to discriminate Xihu longjing tea from Longjing village, Meijiawu village and Geya village and can provide the theoretical basis for the producing region discrimination of Xihu longjing tea.

收稿日期：2015–09–16
基金项目：浙江省重点科技创新团队项目(2010R50028)；“十一五”国家科技支撑计划项目(Y3100246)
作者简介：张龙(1986-)，男，山东临沂人，讲师，主要从事农产品的原产地溯源研究。
E-mail: 10907017@zju.edu.cn

Key words: Xihu longjing tea; near-infrared spectroscopy; spectral pretreatment; principal component analysis; discriminant analysis

0 引言

西湖龙井茶是我国著名的绿茶品牌，居我国名茶之冠，被誉为“绿茶皇后”。由于其极高的历史、文化价值以及优良的品质，深受消费者和市场欢迎。《西湖龙井地理标志》产品规定，只有产自浙江省杭州市西湖区的龙井茶才可以冠上“西湖龙井”品牌商标。由于产量的限制，每年西湖龙井茶都供不应求 [1]。根据 2001 年的《杭州市西湖龙井茶基地保护条例》，西湖龙井茶的产区分为一级产区和二级产区：一级产区指西湖区西湖街道行政区域内的龙井茶基地，主要包括狮峰、龙井、云栖、虎跑和梅家坞，历史上称为“狮龙云虎梅”；二级产区指除一级产区以外的西湖区龙井茶基地 [2]。西湖龙井品牌众多，各品牌的品质和价格各不相同，为规范西湖龙井市场，急需一种不同品牌西湖龙井茶叶的鉴别方法，这对西湖龙井茶的原产地保护具有重要意义。

有机物的近红外光谱是其含氢基团 (X-H) 的合频和倍频，其波数范围在 4000~12500 cm^{-1} 之间。由于近红外光谱的数据容量强大，可将其用于样本中某种或某类有机物的定性与定量检测，已被广泛用于化工和医药产业 [3]。近年来，在食品和农产品的原产地鉴别中，对近红外光谱的研究很多，一些学者对橄榄油、牛肉和小麦等不同物理形态的农产品 (食品) 进行了近红外光谱检测，并结合主成分分析、聚类分析、判别分析、偏最小二乘判别分析和人工神经网络模型等数学方法，建立了这些农产品 (食品) 的原产地鉴别模型 [4−6]。本研究拟通过近红外光谱技术并结合主成分分析和判别分析技术，初步建立 3 个不同地区产的西湖龙井茶的判别方法，为西湖龙井茶的精确快速鉴别提供理论依据。

1 材料与方法

1.1 样品采集及处理

所用西湖龙井茶叶样品采集于 2011 年春季，加工工艺都为手工炒制，茶叶品种都为龙井群体种，共 17 个样品，其中 6 个产自一级产区龙井村 (L)，6 个产自梅家坞 (M)，5 个产自二级产区葛衙庄 (G)。3 组样本产地的地理位置如图 1 所示。3 个地点之间的距离各不相同，龙井村中心和梅家坞中心之间的距离为 2 km，梅家坞到龙井村的中心距离为 3 km，龙井村中心到葛衙庄中心的距离为 5 km。这三个产地的年平均温度为 16.1 °C，平均湿度在 80% 以上，降雨量约为 1500 mm 左右。

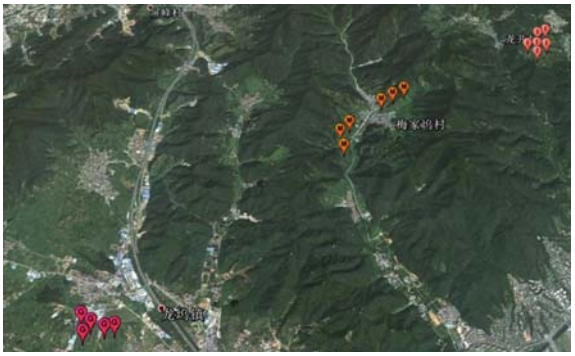

图 1 西湖龙井茶产地的地理位置

将采集的西湖龙井茶样品放在 60 °C 干燥箱中烘干 12 h，然后用四分法各取 10 g 样品放入中药粉碎机中粉碎，将茶叶粉末过 180 目筛，取过筛后的粉末备用。

1.2 近红外光谱采集

用 Nicolet Nexus 870 型傅里叶变换红外光谱仪采集样品的近红外光谱 (图 2)。测量前打开测量室的空调和除湿机，保持温度为 25 °C，相对湿度为 30%；设置近红外光谱仪采集的波数范围为 4000~12000 cm^{-1}，扫描次数为 32 次，分辨率为 2 cm^{-1}。将茶叶粉末平铺在石英杯底部并压实，从石英杯底部采集茶叶的近红外光谱，每个样品重复 3 次，然后取平均值。

图 2　Nicolet Nexus 870 型傅里叶变换红外光谱仪

1.3　数据分析方法

数据分析步骤为：首先，由于近红外光谱含有大量信息，需要将其中的噪音信息去除；接着对光谱进行主成分分析，提取主成分，降低数据的维度和复杂性；然后，用不同类型的判别分析方法进行判别；最后，得到西湖龙井茶的判别模型，通过最优模型筛选模型中输入的主成分数，进一步优化模型。

光谱预处理采用高斯平滑法 (Savitzky-golay Smoothing, SGS)、移动平滑法 (Moving Smoothing, MS)、一阶导数 (First Derivative, 1D)、二阶导数 (Second Derivative, 2D)、一般标准化 (Normalization, Norm)、单位向量标准化 (Unit vector normalization, UVN)、标准正态变换 (Standard Normal Variate, SNV)、标准正态变换和去趋势化法 (Standard Normal Variate and Detrending, SDT)、多元散射校正法 (Multiplicative Signal Correction, MSC) 和中值法 (Center, CT) 等 10 种光谱预处理方法进行。用主成分析 (Principal Component Analysis, PCA) 提取光谱主成分，剔除光谱中的重复元素，降低光谱数据的维度，减少判别模型中输入数据的复杂性。本研究采用线性判别分析 (Linear Discriminant Analysis, LDA)、贝叶斯判别分析 (Bayes Discriminant Analysis, BDA) 和 k 最近邻分析 (k-nearest Neighbor, KNN)3 种判别模型进行分类。光谱预处理、主成分分析和判别分析均采用 Matlab8.3 软件进行。

2　结果与分析

2.1　西湖龙井茶的原始近红外光谱数据

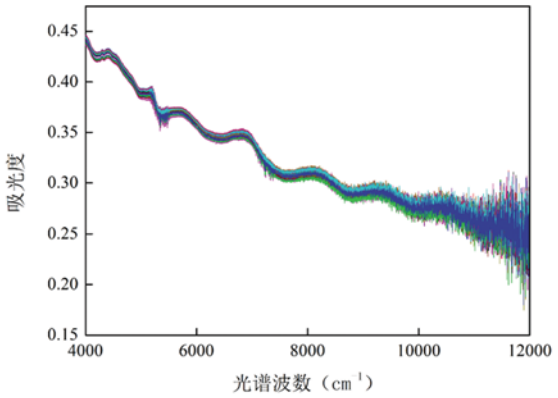

图 3　17 个西湖龙井茶样品的原始近红外光谱图

如图 3 所示，17 个西湖龙井样品的近红外光谱吸收范围在 0.20~0.45 之间，在小波数波段 4000 cm^{-1} 处近红外吸收较高；在大波数波段 12000 cm^{-1} 处的近红外吸收较低，且光谱噪音很大。不同西湖龙井茶样品的近红外光谱在 4432 cm^{-1}、5179 cm^{-1}、5708 cm^{-1}、6805 cm^{-1}、8060 cm^{-1}、9220 cm^{-1} 等处都有明显的吸收峰。

由图 3 可知，不同西湖龙井茶样品的近红外吸收光谱的走势相同，不能用于西湖龙井茶产区的判别。因此，需要通过光谱预处理去除光谱噪音，对光谱进行主成分分析，以光谱主成分得分进行判别分析。

2.2　不同的光谱预处理方法对主成分解释变量的影响

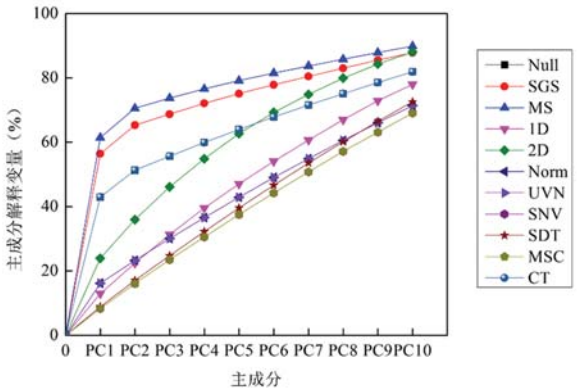

图 4　不同光谱预处理对主成分分析累积方差贡献率的影响

图 4 为采用不同光谱预处理方法对主成分解释变量的影响。原始光谱、SGS 、MS 、CT 和自动变换的前十个主成分解释变量普遍较高，其中第一个主成分解释变量较大在 42.98%~70.66% 之间；前十个主成分解释变量在 81.91%~91.94% 之间。与之相反， 1D 、 Norm 、 UVN 、 SNV 、 SDT 和 MSC 的前十个主成分解释变量较小，在 69.07%~78.00% 之间，其中第一个主成分解释变量更小，在 8.33%~16.18% 之间。 2D 第一个主成分解释变量为 23.92% ，随着主成分数的逐渐增加，解释变量迅速升高，第十个主成分解释变量达 88.04% 。因此，本研究中采用不同光谱预处理方法分析 69.07%~91.94% 范围内的前十个主成分解释变量，接下来将前十个主成分用于判别分析。

2.3 不同光谱预处理方法和判别模型对西湖龙井茶产区判别结果的影响

选用 LDA 、 KNN 和 BDA 共 3 种判别模型对西湖龙井茶产区进行了判别分析。由表 1 可知，原始光谱和 Norm 、 UVN 和 CT 预处理下 LDA 原始判别的正确率为 100% ， MS 和 SNV 预处理后 LDA 原始判别的正确率最小为 82.35% ，原始光谱和不同光谱预处理后 KNN 原始判别的正确率都为 100% ， 1D 、 2D 、 SNV 和 MSC 光谱预处理下 LDA 原始判别的正确率为 100% ，其他预处理下 LDA 原始判别的正确率为 94.12% ；原始光谱和不同光谱预处理后 LDA 交叉验证的判别率为 17.65%~70.59% ， KNN 交叉验证的判别率为 29.41%~41.18% ， BDA 交叉验证的判别率为 29.41%~82.35% 。因此， 2D 光谱预处理后 BDA 原始判别的正确率为 100% ，交叉验证的判别率最高为 82.35% 。所以，选用二阶导数预处理来自不同产区的西湖龙井茶的近红外光谱，选择贝叶斯判别分析判别模型。

表 1 不同光谱预处理对 LDA 、 KNN 和 BDA 判别结果的影响

分类类型	判别类型	Null	SGS	MS	1D	2D	Norm	UVN	SNV	SDT	MSC	CT
原始分类	LDA	100.0	88.24	82.35	88.24	94.12	100.0	100.0	82.35	94.12	94.12	100.0
	KNN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	BDA	94.12	94.12	94.12	100.0	**100.0**	94.12	94.12	100.0	94.12	100.0	94.12
交叉验证分类	LDA	35.29	29.41	29.41	70.59	47.06	17.65	17.65	29.41	17.65	17.65	35.29
	KNN	29.41	41.18	41.18	41.18	47.06	29.41	29.41	29.41	35.29	35.29	29.41
	BDA	29.41	17.65	35.29	76.47	**82.35**	23.53	23.53	35.29	17.65	35.29	29.41

2.4 西湖龙井茶产区的判别模型

在模型筛选过程中，主成分数是一个重要指标。最优模型是用最少的主成分得到最佳的判别结构，因此，进一步研究了主成分数对 2D 光谱预处理正确判别率的影响。如图 5 所示， 1 、 2 、 3 、 4 、 8 个主成分对 BDA 原始判别的结果相同，都是 94.12% ； 5 、 6 、 7 、 9 、 10 个主成分对 BDA 原始判别的结果也都是 100% ； 1~10 个主成分输入 BDA 分析交叉验证判别的正确率为 76.47%~88.24% ；最高交叉验证判别正确率的主成分数为 4 个。为保证判别模型的稳定性，首先选择原始判别结果最优的 5 、 6 、 7 、 9 、 10 个主成分，然后在其中选择交叉验证判别正确率最高的 5 个主成分。

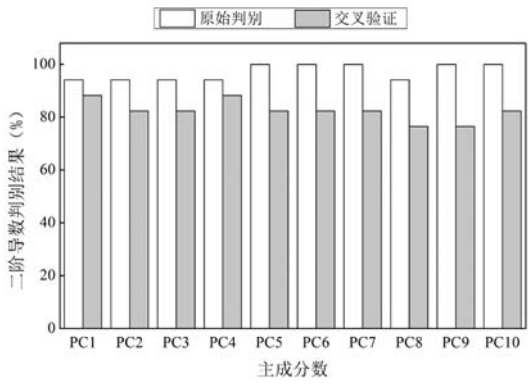

图 5 主成分数对 2D 光谱预处理正确判别率的影响

因此，选用 2D 光谱预处理方法进行噪音去除，然后通过 BDA 进行判别，得到西湖龙井茶的最佳判别模型。原始和交叉验证的正确率分别为 100% 和 82.35%。在交叉验证判别中，葛衙庄、龙井村、梅家坞样品的判别正确率分别为 80%、83.33% 和 83.33%。其中，将 20% 葛衙庄样品错误预测为龙井村样品，将 16.67% 的龙井村样品错误预测为葛衙庄，将 16.67% 的梅家坞样品错误预测为龙井村样品。

表 2a　利用贝叶斯判别分析西湖龙井茶的原始判别率结果

原始判别率	葛衙庄	龙井村	梅家坞	总计
葛衙庄	100.0			100.0
龙井村		100.0		100.0
梅家坞			100.0	100.0

表 2b　利用贝叶斯判别分析西湖龙井茶的交叉验证判别率结果

交叉验证判别率	葛衙庄	龙井村	梅家坞	总计
葛衙庄	80.00	20.00	0	100.0
龙井村	16.67	83.33	0	100.0
梅家坞	0	16.67	83.33	100.0

3　讨论

近年来，近红外光谱技术结合化学计量学的方法已被广泛应用于农产品、食品原产地的溯源 [7]。茶叶中的有机物质包括纤维素、木质素、茶多酚和儿茶素等等。由于土壤、灌溉、农业措施以及加工过程的差别，不同产区的茶叶品质略有差异。这些差异通过人的感官很难区分开，只有通过仪器分析才能鉴别出来。近红外光谱是有机物含氢基团 (X-H) 的合频和倍频。因此，不同地区产的西湖龙井茶因所含有机物不同必然会导致其近红外光谱图不同。这是本研究的理论基础。为去除光谱噪音，采用 10 种光谱预处理方法，经过这些光谱预处理可以不同程度地去除光谱噪音。MSC 可以剔除各样品间由于散射所导致的基线变化 [8]，SNV 能消除因散射和微粒大小引起的多元干涉，1D 和 2D 能消除基线漂移和弯曲，分辨重叠峰，提高灵敏度和分辨率 [9]。如图 2 所示，2D 光谱预处理方法最优，可能是因为 2D 能解析出形状相同的原始光谱的重叠峰，提高灵敏度和分辨率，从而能鉴别出其中化合物的差别。

为获得最佳的判别结果，采用了 3 种判别模型。LDA 是一种有监督的判别分析方法，其判别原理是依据类内方差最小化、类间方差最大化的准则，计算得到判别函数，然后利用此判别函数判别未知样本并进行分类。该方法已被用于统计分析、模式识别以及机器学习中 [10,11]。BDA 根据先验概率分布，首先求出后验概率分布，然后根据各类的后验概率大小进行类别判断 [12]。Cover 等提出了 KNN 算法，其基本原理是在多维空间中找到与未知样本最近邻的 k 个点，并根据这 k 个点的类别来判断未知样本的类，这 k 个点就是未知样本的 k 最近邻 [13−15]。主成分数对判别结果也至关重要，主成分数太少会影响判别结果的正确性，但主成分太多也会产生数据冗余，降低结果的准确性。如图 4 所示，当主成分数达到 7、8 时，贝叶斯判别分析交叉验证的正确判别率反而降低，因此选用了 5 个主成分。

所选的三个地点的地理位置靠近，年平均温度为 16.1 °，平均湿度在 80% 以上，降雨量为 1500 mm 左右。小区域内茶叶受光照、降水等气候因子的影响差异相对较小，我们推测，造成品质不同的因素主要为品种、农艺措施和茶叶加工过程。采集的样品是西湖龙井茶的主栽品种：龙井群体种 [16]。西湖龙井茶区取样点的土壤主要为黄泥土、白砂土、黄筋泥土和油红泥土共 4 种 [17]，土壤类型是造成差异的原因之一。西湖龙井茶的加工采用十大基本炒制手法 “抓、抖、搭、拓、捺、推、扣、甩、磨、压”，炒茶师傅在茶叶炒制过程中用力的大小、炒制的时间都会影响茶叶的品质。据许允文调查，西湖龙井茶栽培管理时，时期不同施肥方法也不同。建国前，主要施用农家土杂肥和菜饼；1960 年开始施有机肥、配方施用氮、磷、钾等化肥。通过进一步检测土壤可以发现，龙井村和梅家坞村茶园土壤的 pH 值和全氮含量都在 4.4 左右，但是梅家坞村

茶园土壤的有机质含量高于龙井村，而龙井村茶园土壤的有效态磷、钾、镁、硫和锌都明显高于梅家坞村 [16] 。因此土壤差异是造成龙井村和梅家坞产西湖龙井茶差异的原因之一。近年来，很多研究者利用近红外光谱对茶叶进行鉴别分析。周健等利用近红外光谱和主成分分析、逐步回归法，建立了不同产地茶叶原料品种的 Fisher 识别模型 [18] 。张龙等通过近红外光谱技术建立了西湖龙井茶与浙江龙井茶的鉴别方法 [7] 。本研究进一步应用二阶导数结合主成分分析和贝叶斯判别分析对西湖龙井茶的三个产区进行鉴别，可为规范西湖龙井茶市场提供理论依据。

4 结论

为鉴别龙井村、梅家坞村和葛衙庄三个地区产的西湖龙井茶，利用近红外光谱技术和数学方法进行了研究。二阶导数光谱预处理、贝叶斯判别分析对龙井村、梅家坞村和葛衙庄三个地区产的西湖龙井茶的鉴别结果最优，判别率在 82.35% 以上。由于本研究中的样品数偏少，以后的研究还要进一步增加样品数以提高判别的正确率。

参考文献

[1] 许咏梅．西湖龙井茶原产地保护实施 10 年后：现状与思考 — 基于龙井村、翁家山、满觉陇的调查 [J].**茶叶**, 2012, **38**(3): 158–164.

[2] 郑旭霞，余继忠，姜新兵，等．西湖龙井茶一级产区施肥现状及建议 [J]. **茶叶**, 2013,**39**(2): 97–100.

[3] Burns D A, Ciurcak EW. Handbook of Near-infrared Analysis [M]. 3rd ed. CRC Press, 2008: 15–16.

[4] 陈永明，林萍，何勇．基于遗传算法的近红外光谱橄榄油产地鉴别方法研究 [J].**光谱学与光谱分析**, 2009, **29**(3): 671–674.

[5] Zalacain A, Ordoudi S A, Díaz-Plaza Eva M, et al. Near-infrared Spectroscopy in Saffron Quality Control: Determination of Chemical Composition and Geographical Origin [J]. *Journal of Agricultural and Food Chemistry*, 2005,**53**(24): 9337–9341.

[6] 张龙，潘家荣，朱诚．近红外光谱分析技术在花生原产地溯源中的应用 [J].**食品科学**, 2013, **34**(6): 167–170.

[7] 张龙，潘家荣，朱诚．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 [J].**红外**, 2013,**33**(3): 44–48.

[8] Fernandez-Cabanas VM, Garridao-Varo A, Perez-marin D, et al. Evaluation of Pretreatment Strategies for Near-infrared Spectroscopy Calibration Development of Unground and Ground Compound Feeding Stuffs [J].*Applied Spectroscopy*, 2006,**60**: 17–23.

[9] Liu L, Philip Y, Saxtonb A M, et al. Pretreatment of Near Infrared Spectral Data in Fast Biomass Analysis [J].*Journal of Near Infrared Spectroscopy*, 2010,**18**(5): 317–331.

[10] 张新新，李雨，纪玉佳，等．主成分－线性判别分析在中药药性识别中的应用 [J].**山东大学学报 (医学版)**, 2012, 50(1): 143–146.

[11] Thomaz C E, Kitani E C. A Maximum Uncertainty LDA Based Approach for Limited Sample Size Problems-with Application to Face Recognition [J].*Journal of the Brazilian Computer Society*, 2006, **12**(2): 7–18.

[12] 肖培，崔步云．贝叶斯判别分析在布氏杆菌常见种别鉴定中的应用 [J].**中国卫生统计**, 2013 ,**30**(6): 802–804.

[13] 许东，代力民，邵国凡，等．基于 RS 、 GIS 及 k-近邻法的森林蓄积量估测 [J].**辽宁工程技术大学学报 (自然科学版)**, 2008,**27**(2): 195–197.

[14] Cover T M, Hart P E. Nearest Neighbor Pattern Classification [J].*IEEE Transactions on Information Theory*, 1967, **13**(1): 21–27.

[15] Henley WE ， Hand DJ. A k-nearest-neighbor Classification for Assessing Consumer Credit Risk[J].*The Statistician*, 1996,**44**: 77–95.

[16] 邵银泽，徐德玉．全面实施原产地保护再创国饮龙头之辉煌 [J]. **杭州农业科技**, 2006(3): 2–4.

[17] 方晨，邹新武．浅析西湖龙井茶品质形成的原因 [J].**中国茶叶加工**, 2011, **22** (3): 27–30.

[18] 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.