基于机器学习的 NIR 光谱柑橘产地鉴别框架

但松健
(重庆第二师范学院 继续教育学院,重庆 400067)

摘要:研究基于机器学习的近红外(Near infrared,NIR)光谱柑橘产地鉴别模型,根据 NIR 光谱数据的特点,提出了一个完整的产地鉴别通用框架,包括数据预处理、特征选择、模型建立和交叉验证等步骤。在框架下对比多种预处理算法以及多种机器学习算法,基于 NIR 光谱进行柑橘产地鉴别,得到了较好的识别结果,提高了柑橘产地鉴别的准确性。

关键词:机器学习; NIR 光谱; 产地鉴别
中图分类号:TP81 **文献标识码**:A **文章编号**:1008 - 6390(2019)04 - 0117 - 06

近红外光谱分析技术作为一种快速、准确、便捷且非破坏性的分析技术,在农产品品质检测和产地鉴别方面得到了广泛应用,被认为是有望替代传统化学分析的无损检测方法[1-4]。目前,基于近红外光谱分析的柑橘产地鉴别技术还较为耗时费力且不够精确,其完整性、系统性和操作性还与实际应用有很大差距,建立一套能对柑橘产地进行快速鉴别的有效技术体系,对于柑橘产业在我国的健康发展有着重要的作用[5-6]。

一、基于机器学习的 NIR 光谱柑橘产地鉴别框架

本文通过基于机器学习的光谱分析技术建立了一种快速无损的柑橘产地鉴别通用框架,具体流程如图 1 所示。首先,采用预处理算法对光谱进行整形降噪,从而降低原始数据中的噪声对分类器的干扰;其次,采用 PCA 方法对降噪后的 NIR 光谱进行特征抽取,从而将高维数据降维到适当的维度;然后,利用特征选择算法对降维后的光谱数据进行适当的特征选择以利于分类器更快更精确地学习;最后,选择不同的分类器,在统一的训练框架和性能评价指标下,选出最优的分类器建立光谱识别模型[7-12]。

二、实验结果及分析

在实验中,选取了常见的朴素贝叶斯、最近邻分类(KNN)以及决策树算法作为测试分类器[13-14],对采集的 6 个省市 16 个不同地区的柑橘进行产地鉴别。原始近红外光谱的范围为 1000 ~ 2499 nm,原始特征维度为 1500 维。每个地区约采集 100 个柑橘样本,总的样本数量为 1558 个。根据鉴别框架对原始光谱数据进行预处理、特征抽取、特征选择以及模型交叉验证,以得到最后的性能评价。所有的模拟实验都在 Windows 7 平台使用 Matlab 2008b 实现,使用了统计工具箱和数据挖掘工具箱。

(一)原始光谱及预处理结果

考虑到近红外光谱仪器、实验环境和操作误差带来的不可避免的噪声,对原始数据进行预处理以去除噪声干扰是非常必要的。采用 SG 平滑法对光谱进行整形,SG 平滑在 121 大小的窗口下进行,并用到了原始 SG 平滑及在此基础上衍生出的一阶和二阶导数。这三种去噪方法以及原始光谱的信息如图 2 所示。

由图 2 可以看出,经过 SG 平滑,原始光谱图变得平滑。在进行一阶导数运算后,光谱范围从[0,1]压缩到[-0.002,0.006],光谱信号进一步平滑。从二阶导数的结果看,平滑效果跟一阶导数接近,但数据得到进一步压缩,范围缩小到[-0.00009,0.00007]。虽然导数操作可以进一步平滑数据,但也可能会丢失部分具有区分度的细节。因此,去噪

收稿日期:2019 - 03 - 14
基金项目:重庆第二师范学院校级课题"基于近红外光谱分析的柑橘产地鉴别技术研究"(KY201711B)
作者简介:但松健,博士,副教授,研究方向:机器学习和数据挖掘。

预处理操作需要进行合适的选择。通过图 2 可以看出,16 个地区柑橘样本的光谱具有很大的重叠性,如果直接使用这些数据(1500 维)进行识别具有很大的挑战性。

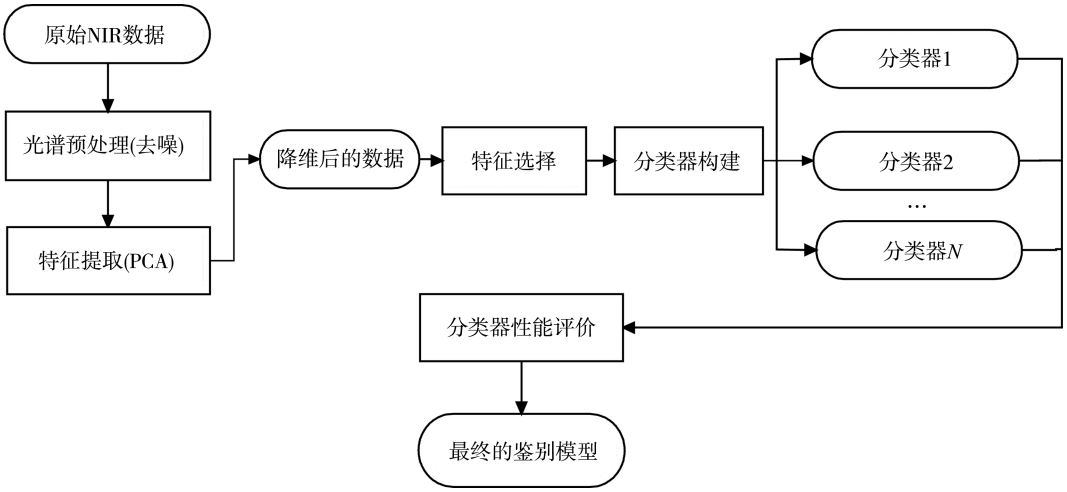
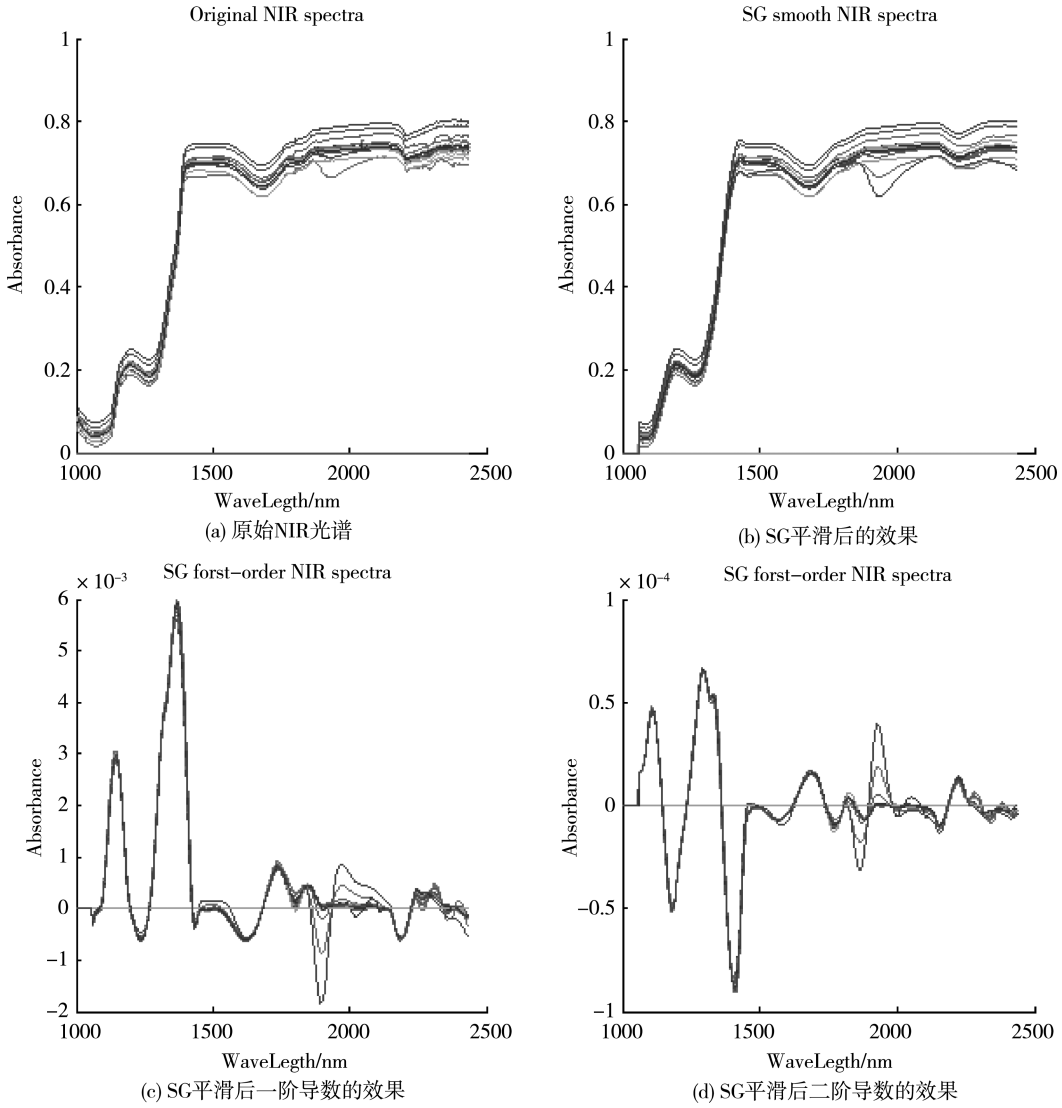


图 1　基于机器学习的 NIR 光谱产地鉴别框架



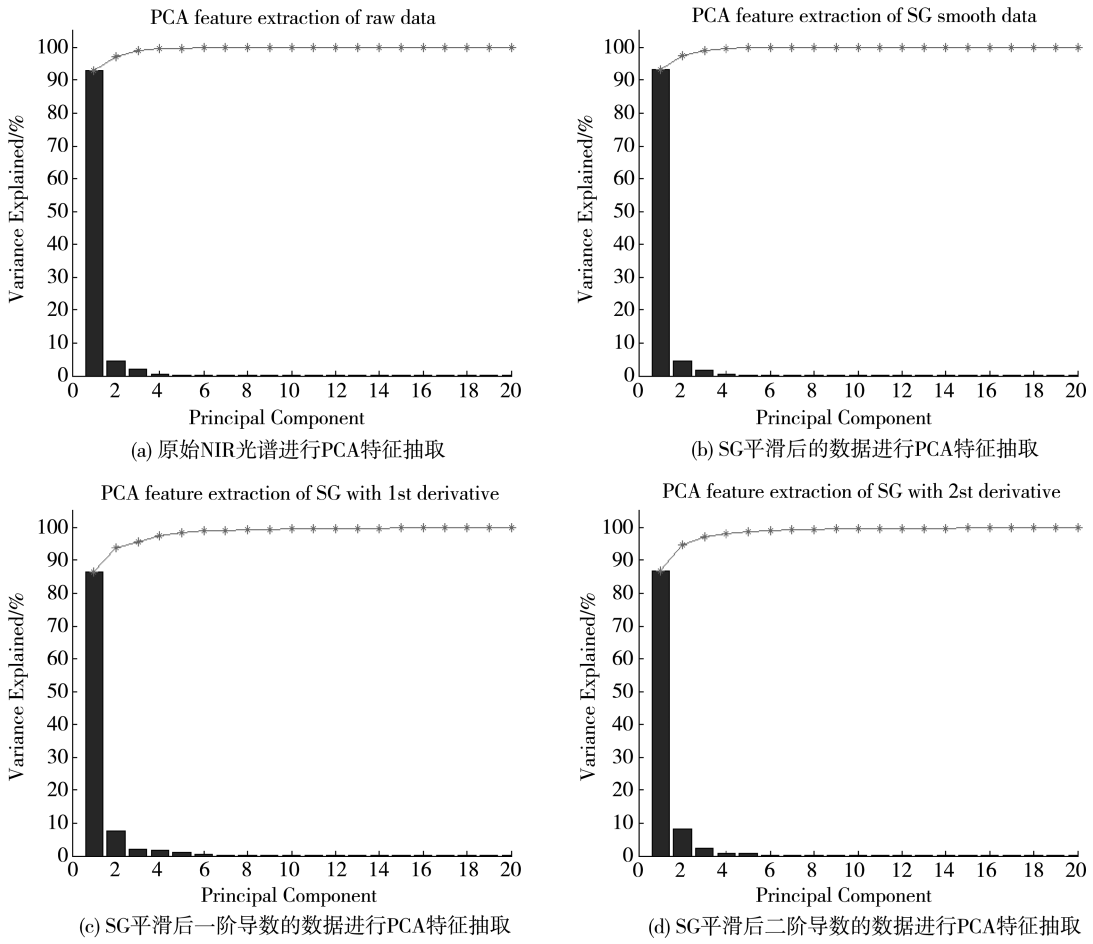
(a) 原始NIR光谱　(b) SG平滑后的效果

(c) SG平滑后一阶导数的效果　(d) SG平滑后二阶导数的效果

图 2　柑橘原始光谱及去噪后效果

(二)特征抽取结果

从上一小节的实验中可以看出,经过去噪的数据并不适合用分类器进行直接训练,需要进行适当的特征抽取,以便提取主要信息,去除不必要的冗余信息,在识别框架下采用 PCA 方法来提取光谱的主成分。因为没有足够的证据表明某一段光谱具有很强的区分度,因此对整个光谱段(1000 ~ 2499 nm)进行主成分提取以得到最具代表性的光谱信息,以主成分的贡献度排序得到的结果如图 3 所示。



(a) 原始NIR光谱进行PCA特征抽取　(b) SG平滑后的数据进行PCA特征抽取

(c) SG平滑后一阶导数的数据进行PCA特征抽取　(d) SG平滑后二阶导数的数据进行PCA特征抽取

图 3　柑橘 NIR 光谱数据进行 PCA 特征抽取之后的主成分贡献度

一般来说,建立模型所需要的主成分个数往往由前几个最有代表性的主成分所占光谱信息的比重来决定。如图 3 所示,柱状图代表该主成分的贡献度(即所含信息在整个数据集中的比重),红色的点代表其前 N 个主成分累积贡献度。从图 3 中可以看出,前 3 个主成分占据了很大的比重,例如在图 3(a)中,对原始的光谱数据进行 PCA 降维,前 3 个主成分占据了 98.98% 的信息量。对 SG 平滑后的数据提取主成分,前 3 个主成分占据了 99.11% 的信息量,而对一阶和二阶导数后的平滑数据,前 3 个主成分分别占据了 95.17% 和 97.16% 。

虽然前 3 个主成分能有效表示之前的原始数据集,但对于分类器来说,其代表的信息或许并不具有区分度。例如,对原始数据和采用不同平滑算法的前两个主成分的联合分布情况,用散点图来表示,如图 4 所示。为了更好地显示其分布特性,这里只画出了 20 个来自 5 个不同地区的柑橘光谱样本,包括四川武胜,浙江临海,重庆巫山、奉节和北碚。

从图4 可以看出,在原始光谱和 SG 平滑后的光谱数据上进行 PCA 降维后,不同省市之间的 PC 分布具有一定的区分度,而位于重庆的 3 个不同产地的样本由于采集区域较近,柑橘生长环境较为类似,因此出现了一定程度的重叠。使用 SG 平滑结合一阶和二阶导数法后,样本的分布空间被扩展,从而加大了样本间的分散度,但也进一步增加了样本重叠的区域。无论采用哪种方法,柑橘样本的前两个 PC 直接进行识别都存在着一定的难度。因此,可以适当加入更多的 PC 特征增加其辨识度,我们取前 20 个 PC 作为训练特征输入分类器中。

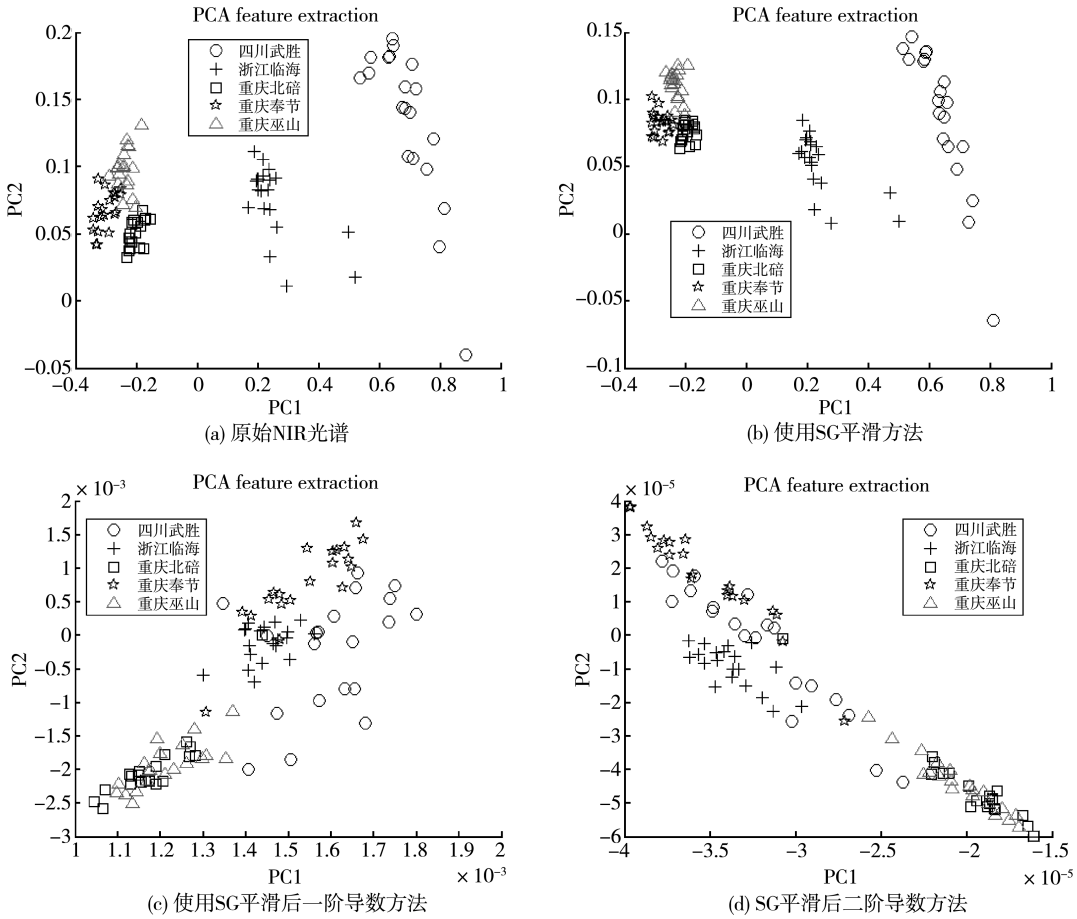


图 4　5 个地区的柑橘 NIR 做 PCA 特征抽取后，贡献度第一和第二的主成分分布

（三）特征选择及分类器性能结果

通过数据平滑和主成分提取后，主要采用了机器学习算法中的常见分类器，包括了决策树算法（DT）、贝叶斯分类器（NB）、K 近邻分类器（KNN）和线性判别分类器（LDA），对 6 个省市共计 16 个地区的柑橘样本进行了产地鉴别模型的建立。根据提出的产地鉴别框架，所有的分类器都进行了 5 × 10 次交叉验证，并将 50 次运行后的平均识别率作为输出结果，各个分类器性能如表 1 所示。

首先，在没有进行特征选择的情况下，表 1 统计了测试的 4 个分类器平均准确率 P_a。

表 1　无特征选择时，测试的 4 个分类器的产地鉴别平均准确率 P_a　%

分类器	原始数据	SG 平滑	SG 平滑 1 阶导数	SG 平滑 2 阶导数
DT	83.5	85.1	79.0	69.4
KNN	88.2	89.3	86.6	76.6
NB	88.8	91.3	88.2	81.3
LDA	92.6	92.5	92.6	86.7

注：DT 为决策树，NB 为贝叶斯，KNN 为最近邻，LDA 为线性判别

从表 1 可以看出，LDA 分类器在各个数据集上的表现最优，最高达到了 92.6% 的平均准确率，其次是 KNN 和 NB 分类器。在数据平滑算法方面，相比原始数据集，在采用 SG 平滑的数据集上，DT、NB 和 KNN 分类器的性能都得到了明显提高，而 LDA 算法变化不大，但 SG 平滑后结合导数的方法反而降低了识别精度，特别是导数阶数越多、效果越差，其原因可能是过多的平滑导致了具有区分度的特征的丢失。

为了进一步显示交叉验证中 50 次测试的分类器的性能及其稳定性，通过 4 个分类器在不同平滑算法下的盒图[14]发现，使用 SG 平滑后大部分分类器的预测准确率达到了最高（除 LDA 与采用原始数据持平），并且最为稳定，而采用一阶和二阶导数后，由于数据被过度平滑，影响了其稳定性。

除了准确率，本文还统计了其他性能指标，如敏感度（TPR）、特异性（FPR）和综合指标 F_1，结果如表 2 所示。

表 2 结果与表 1 类似，在各项性能指标上，LDA 仍然得到了最高的识别率，DT、KNN 和 NB 分类器在 SG 平滑的数据集上识别结果较好。

表 2　无特征选择时，测试的 4 个分类器的产地鉴别平均敏感度、特异性和综合指标 F_1 值

数据集	评价指标	DT	KNN	NB	LDA
原始数据	*TPR*	78.5	83.3	83.9	87.0
	FPR	78.5	82.6	83.6	87.0
	F_1	78.3	82.8	83.6	87.0
SG 平滑	*TPR*	79.9	84.3	85.9	86.9
	FPR	79.9	83.7	85.9	87.0
	F_1	79.7	83.9	85.7	86.9
SG 平滑 1 阶导数	*TPR*	75.0	81.9	83.6	87.0
	FPR	74.3	81.5	83.0	87.0
	F_1	74.4	81.5	83.0	86.9
SG 平滑 2 阶导数	*TPR*	66.3	74.2	77.3	82.2
	FPR	65.8	72.2	77.0	81.8
	F_1	65.6	72.8	76.9	81.8

注：DT 为决策树，NB 为贝叶斯器，KNN 为最近邻，LDA 为线性判别

对 PCA 降维后的特征进行进一步的选择，对同样的分类器和数据集进行了交叉验证，结果如表 3 所示。经过特征选择后，LDA 模型依旧获得了最高的识别准确度，但相比特征选择前的提高并不明显，原因在于 LDA 在寻求最佳的投影方向时已经考虑具有最大区分度的特征投影方向，而其他模型相比特征选择前的性能都有了明显的提高，KNN 和 NB 都达到了较高的识别度（≥90%），特别是在采用二阶导数法平滑的数据集上，测试的 4 个分类器都有了较大的提升。提高最多的为 DT 和 KNN 模型，平均准确率分别从 69.4% 和 76.6% 提高到了 80.4% 和 88.0%。

表 3　进行特征选择后，测试的 4 个分类器的产地鉴别平均准确率 P_a

分类器	原始数据		SG 平滑		SG 平滑 1 阶导数		SG 平滑 2 阶导数	
DT	88.8	+5.3	89.1	+4.0	86.3	+7.3	80.4	+11.0
KNN	91.5	+3.4	91.5	+2.1	90.4	+3.8	88.0	+11.4
NB	91.3	+2.5	92.0	+0.8	91.7	+3.5	90.2	+9.0
LDA	92.5	−0.1	92.7	+0.2	92.8	+0.2	92.4	+5.6

注：右侧数据为对比未进行特征选择的分类器的结果差异，“+”号表示较之前有所提升，“−”表示识别率下降

最后，表 4 给出了进行特征选择后，基于敏感度（TPR）、特异性（FPR）和综合指标 F_1 的结果。可以看出，在进行特征选择后，KNN 和 NB 达到了与 LDA 相近的性能，DT 模型的识别效果也有显著提升，而 LDA 提升不大，并且各个数据集的性能差异并不明显。

表 4　进行特征选择后，测试的 4 个分类器的产地鉴别平均敏感度、特异性和综合指标 F_1 值

数据集	评价指标	DT	KNN	NB	LDA
原始数据	*TPR*	83.6	86.1	85.8	86.8
	FPR	83.4	85.7	85.8	86.9
	F_1	83.4	85.8	85.8	86.8
SG 平滑	*TPR*	83.9	86.0	86.6	87.1
	FPR	83.7	85.6	86.6	87.1
	F_1	83.7	85.7	86.4	87.1

数据集	评价指标	DT	KNN	NB	LDA
SG 平滑 1 阶导数	*TPR*	81.3	85.2	86.1	87.0
	FPR	81.0	85.1	86.3	87.2
	F_1	81.0	85.0	86.1	86.9
SG 平滑 2 阶导数	*TPR*	76.1	83.2	84.9	86.8
	FPR	75.8	82.8	84.9	86.8
	F_1	75.8	82.9	84.8	86.8

注:DT 为决策树,NB 为贝叶斯,KNN 为最近邻,LDA 为线性判别

三、结语

本文针对柑橘光谱产地识别问题,提出了一个通用识别框架并在该框架下对柑橘样本进行了产地鉴别。首先,采用 SG 平滑法以及 SG 平滑结合一阶和二阶导数法对数据进行平滑,并采用 PCA 对数据降维以抽取最有代表性的特征,之后利用特征选择算法对抽取后的特征进行最有区分度的选择,最后采用决策树、最近邻、朴素贝叶斯和线性判别分析模型,对 16 个地区的柑橘数据建立产地鉴别模型。实验结果表明,SG 平滑算法能增强大部分分类器的识别能力,特征选择算法也对柑橘产地的鉴别有积极作用。在测试的分类器中,LDA 的性能最为稳定,并获得了最优的产地鉴别准确率 92.8%。

参考文献:

[1]刘姗姗,吴志生,邢玲,等.基于显微近红外光谱技术的天然牛黄和人工牛黄的鉴别研究[J].中华中医药杂志,2014(1):84-87.

[2]黄艳华,杜娟,夏田,等.近红外光谱在植物种及品种鉴定中的应用[J].中国农学通报,2014(6):46-51.

[3]Caligiani A, Palla L, Acquotti D, etal. Application of 1 H NMR for the characterisation of cocoa beans of different geographical origins and fermentation levels[J]. Food Chemistry, 2014(157):94-99.

[4]Sádecká J, Jakubíková M, Májek P, et al. Classification of plum spirit drinks by synchronous fluorescence spectroscopy [J]. Food Chemistry, 2016(196):783-790.

[5]Li G, Nunes L, Wang Y, et al. Profiling the ionome of rice and its use in discriminating geographical origins at the regional scale, China[J]. Journal of Environmental Sciences, 2013, 25(1):144-154.

[6]Caligiani A, Palla L, Acquotti D, et al. Application of 1 H NMR for the characterisation of cocoa beans of different geographical origins and fermentation levels[J]. Food Chemistry, 2014(157):94-99.

[7]Chen H, Lin Z, Wu H, et al. Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 135(0):185-191.

[8]Diniz P H G D, Gomes A A, Pistonesi M F, et al. Simultaneous classification of teas according to their varieties and geographical origins by using NIR spectroscopy and SPA-LDA[J]. Food Analytical Methods, 2014, 7(8):1712-1718.

[9]Ren G, Wang S, Ning J, et al. Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS)[J]. Food Research International, 2013, 53(2):822-826.

[10]Zhao H, Guo B, Wei Y, et al. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat[J]. Food Chemistry, 2013, 138(s 2/3):1902-1907.

[11]Shen T, Zou X, Shi J, et al. Determination geographical origin and flavonoids content of Goji Berry using near-infrared spectroscopy and chemometrics[J]. Food Analytical Methods, 2015, 9(1):1-12.

[12]Wang J J, Yan S M, Yang B. Determination of ginsenosides amount and geographical origins of ginseng by NIR spectroscopy[J]. Spectroscopy and Spectral Aralysis. 2015, 35(7):1885-1888.

[13]Asir D, Appavu S, Jebamalar E. Literature review on feature selection methods for high-dimensional data[J]. International Journal of Computer Applications, 2016(1):9-17.

[14]El-bendary N, El-hariri E, Hassanien A E, et al. Using machine learning techniques for evaluating tomato ripeness [J]. Expert Systems with Applications, 2015, 42(4):1892-1905.

[责任编辑 乡 下]