油脂安全

DOI: 10.19902/j.cnki.zgyz.1003-7969.220236

基于数据融合策略植物油光谱模式的识别

邱薇纶[1],周燕舞[2],石孟良[2]

(1.湖南警察学院 刑事科学技术学系,长沙 410138; 2.湖南省湘潭县公安局刑侦大队,湖南 湘潭 411228)

摘要:为实现对不同植物油的快速无损分类识别,探究数据融合技术在提升光谱分类模型精度方面的可行性与应用价值,借助衰减全反射-傅里叶变换红外光谱技术、表面增强拉曼光谱技术结合多源数据融合技术,开展了对7种共计180份植物油样本的分类识别。基于单一光谱模型、数据层融合模型和特征层融合模型,比较了Bayes判别分析(BDA)和多层感知器神经网络(MLP)两种化学计量学方法在区分各样本时的差异,同时考察了主成分分析、广义最小平方、最大似然、主轴因式分解4种算法在特征提取方面的差异。结果表明,光谱数据融合在识别植物油方面具有显著的优势,BDA模型对各样本的区分能力强于MLP模型,相较于其他3种算法,主成分分析在油样特征提取方面展现了较为理想的结果。基于PCA特征提取的特征层融合BDA模型为最佳识别模型,以此实现了180份植物油样本100%的准确区分,同时对5种品牌花生油达到了100%的准确区分,实现了对各样本"种类-品牌"的两级识别分类工作。

关键词:植物油;光谱;数据融合;特征提取;识别

中图分类号:TS227;O657 **文献标识码**:A **文章编号**:1003-7969(2023)07-0062-06

Spectral pattern recognition of vegetable oils based on data fusion strategy

QIU Weilun[1], ZHOU Yanwu[2], SHI Mengliang[2]

(1. School of Forensic Science, Hunan Police College, Changsha 410138, China; 2. Criminal Investigation Brigade of Xiangtan County Public Security Bureau, Xiangtan 411228, Hunan, China)

Abstract: In order to achieve the rapid and non-destructive recognition of different vegetable oils, and explore the feasibility and application value of data fusion technology in improving the classification accuracy of spectral models, a total of 180 vegetable oil samples from 7 kinds were recognized and classified by attenuated total reflectance-Fourier transform infrared spectroscopy, surface-enhanced Raman spectroscopy and multi-source data fusion. The differences of Bayes discriminant analysis (BDA) and multilayer perceptron neural network (MLP) in classifying all samples were compared and discussed based on single model, data layer fusion model and feature layer fusion model. Besides, the differences of principal component analysis, generalized least square, maximum likelihood and principal axis factorization in feature extraction were investigated. The results showed that the spectral data fusion had significant advantages in recognizing vegetable oils. The ability of BDA model to distinguish each sample was more predominant than that of MLP. Compared with another three algorithms, principal component analysis showed the more positive results in extracting features. The BDA model based on feature layer fusion from PCA feature extraction was considered as the optimal model, and it could achieve 100% accurate differentiation of 180 samples and 100% accurate differentiation of 5 brands of peanut oil, realizing the two-level recognition and classification of samples from kind to brand.

Key words: vegetable oil; spectral; data fusion; feature extraction; recognition

收稿日期:2022-04-06;**修回日期**:2023-03-02

基金项目:湖南省自然科学基金面上项目(2023JJ30221)

作者简介:邱薇纶(1982),女,讲师,硕士,主要从事刑事技术方面的研究(E-mail)915177230@qq.com。

植物油在人们生活生产中扮演着十分重要的角色,膳食烹饪、工业生产、燃料消耗都离不开它,它与人们的日常息息相关。侦查员往往会在案发现场和

嫌疑人衣物上发现并提取到相关的油样物证。检验和识别这些物证,一定程度上有助于辨别嫌疑人是否做如实供述,刻画和推断其饮食习惯与职业特点。

目前,对植物油的检验主要集中在营养成分[1]、添加剂[2-3]、掺伪鉴别[4-5]等方面。植物油检验在司法鉴定中的应用也有一定报道[6],但相对较少,这些方法大多对物证有损坏[6-7],油样物证较少时无法开展二次复检工作,而且所用方法相对烦琐,操作要求较高,不便于侦查人员开展快速筛查工作。因此,建立简便易行、快速无损的植物油检验方法显得尤为必要。

分子光谱分析技术作为经典的无损检验方法,有着较为广泛的应用,结合化学计量学方法开展对物质光谱信息的挖掘和解读已然是当下研究的热点和重点之一[8-10]。光谱数据融合可以结合不同来源的光谱信息,基于不同融合级别方式,借助化学计量学等相关方法构建模型,从而实现更为全面反映样品潜在信息价值的目的,达到不同光谱技术间的优势互补。目前尚未有报道将其应用于植物油的检验研究之中。红外光谱技术和拉曼光谱技术在分子基团定性检测中有良好的优势互补特性,在物质的定性分析中将二者融合具有明显的优势[11]。根据融合方式不同,分为数据层融合与特征层融合。在融合模型中,不同融合方式各有利弊。数据层融合方式简单方便,可直接对原始的样本数据进行融合并开展分析;但是,当数据维度较高时,其计算时间和计算复杂度均不占据优势。特征层融合方式可以解决这一不足,它能通过特征提取来降低计算的时间和复杂度,然而其复杂程度要比数据层融合方式高。

鉴于此,本实验借助衰减全反射-傅里叶变换红外光谱(ATR-FTIR)和表面增强拉曼光谱(SERS)分析技术获取市面常见植物油样本光谱信息数据,基于不同融合级别方式,借助 Bayes 判别分析和多层感知器神经网络两种化学计量学方法构建植物油分类识别模型,以期实现对不同植物油的准确识别和分类,为快速无损检验的应用和发展提供一定的参考和借鉴。

1　材料与方法

1.1　实验材料

结合实际案件情况,从市面上收集常见的 7 种植物油样本共计 180 份,180 份油样的基本信息见表 1。

表 1　180 份样本的基本信息

植物油	品牌	数量(份)
橄榄油	克鲁托	30
花生油	鲁花	7
花生油	胡姬花	8
花生油	金龙鱼	6
花生油	福临门	8
花生油	刀唛	11
葵花籽油	多力	20
山茶油	绿海	10
椰子油	娜古香	10
玉米油	长寿花	20
芝麻油	燕庄	12
芝麻油	古币	20
芝麻油	老榨坊	18

Nicolet is10 型傅里叶变换红外光谱仪、Nicolet Almega XR 型拉曼光谱仪,Thermo Scientific 公司;Easypeak 拉曼增强试剂(银溶胶),上海纳腾仪器公司。

1.2　实验方法

1.2.1　光谱测定

傅里叶变换红外光谱条件:扫描次数 64 次,光谱分辨率 2 cm^{-1},测量范围 650 ~ 4 500 cm^{-1},动态调整 130 000 次/s,信噪比 50 000∶1,FTIR 标准线性度(ASTM E1421)<0.1% T,峰-峰噪声值(DTGS 检测器,KBr 窗片)<1.24×10^{-5}AU。

拉曼光谱条件:拉曼增强基底探针分子 2-巯基吡啶,拉曼增强基底检测限 3×10^{-7} mol/L,空间分辨率 1 μm,共聚焦深度剖析分辨率 2 μm,光谱分辨率 2 cm^{-1},激光光源 780 nm,扫描时间 8 s,曝光次数 4,光谱测量范围 200 ~ 3 000 cm^{-1}。

1.2.2　光谱数据预处理

在光谱测量的过程中,往往由于仪器自身原因、光源条件、实验温度等影响,获取的样本光谱存在基线漂移、高频噪声等现象[12],对所获取的样本光谱数据进行预处理可有效消除这些不利因素带来的负面影响。借助自动基线校正、峰面积归一化、多元散射校正、Savitzky-Golay 平滑对原始光谱数据进行预处理,以消除噪声和干扰数据。

1.2.3　Bayes 判别分析

Bayes 判别分析(Bayes discriminant analysis,BDA)是一种较为经典的分类方法,其主要是在考虑不同类别样本先验概率的前提下,按照一定准则构造判别函数,分别计算单个样本落入各个类别的概率,所得结果最大的那一类即为该样本所属的类别。其基本原理是:假设 n 为总体样本数,n_j 为第 j

个总体的样本数，μ_j 为第 j 个总体的均值向量，则 Bayes 判别函数为：

$$f(j/x) = \ln q_i + C_{(j)} + C_{j(j)}\, x_i \quad (1)$$

式中：q_i 为先验概率，$q_i = n_j/n$；$C_{(j)} = -\left(\frac{1}{2}\right)(\mu_{(j)})^T \sum_{(j)}^{-1}\mu_{(j)}$，$\sum_{(j)}^{-1}\mu_{(j)}$ 为第 j 个总体协方差矩阵[13-14]；$C_{j(j)}\, x_i = \sum_{(j)}^{-1}\mu_{(j)}$。

1.2.4 多层感知器神经网络

多层感知器神经网络（Multilayer perceptron neural network，MLP）是一种含有多个隐藏层同时具有映射特性（即由输入层向输出层传输）的分类算法，其具有较深层次的网络结构，加入非线性因素，保证了隐藏层在充分组合特征方面的优越性，由于不再是简单的线性组合，这使得神经网络表达能力变得更加强大，几乎可以逼近任意函数，这在特征提取和分类识别方面尤为有效[15-16]。假设输入层用向量 $\boldsymbol{X}$ 表示，则隐藏层的输出为 $y(x) = f(W_1\boldsymbol{X} + b_1)$。式中：$W_1$ 为权重；b_1 为常数；函数 f 为常用的 sigmoid 函数或者 tanh 函数。函数 f 的输出层是：

$$f(a) = \frac{e^a}{\sum_{k=1}^{N} e^{a_k}} \quad (2)$$

式中：N 为输出层中节点的数量。

1.2.5 单一光谱模型的建立

对预处理的光谱数据借助主成分分析（Principal component analysis，PCA）提取特征变量，以此构建 BDA 和 MLP 分类模型。

1.2.6 数据层融合模型的建立

数据层融合策略是指将采集到的原始光谱数据直接进行拼接融合，基于新的数据集构建分类识别模型。新的光谱矩阵所囊括的波段信息较多，模型中可能会包含较多的无关信息等，这可能会降低模型的准确度，增加模型的运行时间[17]。基于此，将各样本经预处理的红外光谱与拉曼光谱信息数据拼接融合，采用 PCA 提取特征变量，分别构建 BDA 和 MLP 分类识别模型。

1.2.7 特征层融合模型的建立

特征层融合策略先对原始光谱数据进行特征提取，而后对特征信息进行综合分析和处理，从而构建具有较强区分能力的分类模型。本实验采用 PCA、广义最小平方（Generalized least square，GLSE）、最大似然（Maximum likelihood，ML）、主轴因式分解（Principal axis factorization，PAF）4 种算法分别提取各样本红外光谱与拉曼光谱数据特征，而后将其融合并建立 BDA 和 MLP 分类模型。

2 结果与讨论

2.1 油样光谱数据的预处理

经预处理后 180 份植物油样本的光谱图如图 1 所示。

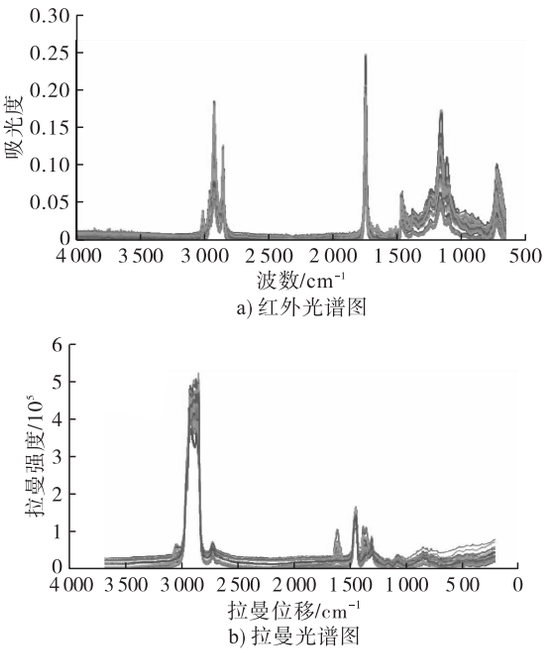


图 1 预处理后 180 份植物油样本的光谱图

2.2 单一光谱模型识别分类结果

单一光谱模型对各样本的总体识别分类结果如图 2 所示。

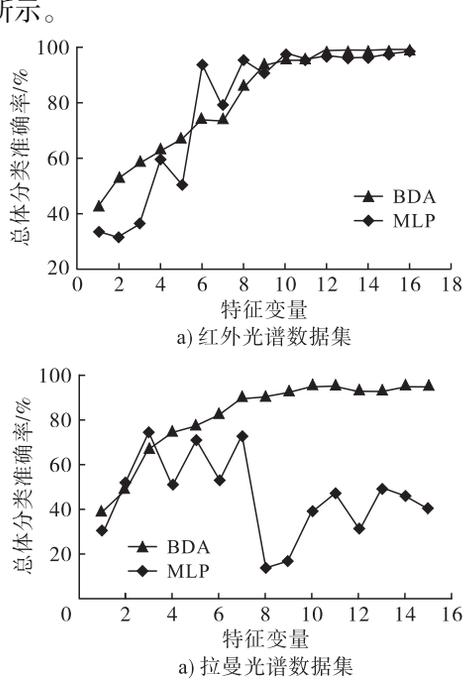


图 2 单一光谱模型对各样本的总体识别分类结果

由图 2a 可知：红外光谱数据集模型中，特征变量取前 12 时，BDA 模型对各样本分类准确率最高（98.9%）；特征变量取前 16 时，MLP 模型对各样本

的分类准确率最高(98.6%),特征变量取前 6、7、8、10 时,MLP 模型的精度相对高于 BDA 模型,但总体来看,BDA 模型对各样本的区分能力略强于 MLP 模型。由图 2b 可知:拉曼光谱数据集模型中,特征变量取前 10 时,BDA 模型对各样本分类准确率最高(94.9%);特征变量取前 3 时,MLP 模型对各样本的分类准确率最高(74.2%),特征变量取前 2、3 时,MLP 模型的精度相对高于 BDA 模型,但总体而言,BDA 模型对各样本的区分能力仍强于 MLP 模型。分析认为,BDA 模型的分类结果是一个基于训练数据和先验概率的模型参数的分布,它在理论上具有最小的误差率,加之它对缺失数据不太敏感,这就保证了它在多分类问题上的优越性[16-18]。

综上,虽然随着特征变量的增多,单一光谱模型对各样本的识别能力增强,但未能实现 100% 的准确区分。

2.3　数据层融合模型识别分类结果

数据层融合模型对各样本的总体识别分类结果如图 3 所示。

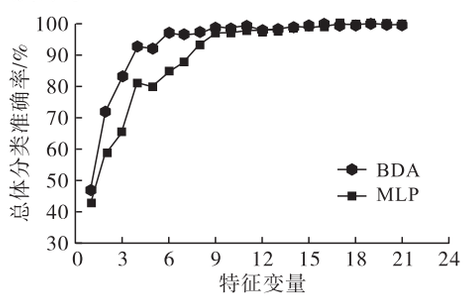


图 3　数据层融合模型对各样本的总体识别分类结果

由图 3 可知,随着特征变量的增多,模型对各样本的识别能力增强。特征变量取前 16、17 时,BDA 和 MLP 分类模型对各样本均实现了 100% 的准确分类。相比较单一光谱模型,数据层融合模型在一定程度上提升了对各样本的分类能力,原因在于红外光谱和拉曼光谱的优势互补,弥补了单一来源分析信号信息量不足的缺点[18],实现了模型对油样光谱信息更为全面的解读和挖掘。这表明光谱融合策略在开展油样模式识别分类工作中是十分有必要的,数据层融合方式可以实现对油样的较为准确分类。此外,BDA 模型对各样本的识别分类能力相对强于 MLP 模型。

2.4　特征层融合模型识别分类结果

特征层融合模型对各样本的总体识别分类结果如图 4 所示。

由图 4 可知,随着特征变量的增多,模型对各样本的识别能力增强。在基于 PCA 构建的分类模型中,特征变量分别取前 15 和 18 时,BDA 和 MLP 分类识别模型实现对各样本 100% 的准确分类。在基于 GLSE 构建的分类模型中,特征变量均取前 18 时,BDA 和 MLP 分类识别模型实现对各样本 100% 和 99.4% 的准确分类。在基于 ML 构建的分类模型中,特征变量分别取前 20 和 15 时,BDA 和 MLP 分类识别模型均分别实现对各样本 100% 和 97% 的准确分类。在基于 PAF 构建的分类模型中,特征变量分别取前 20 和 17 时,BDA 和 MLP 分类识别模型分别实现对各样本 100% 和 99.4% 的准确分类。相比较于其他 3 种算法,PCA 在油样特征提取方面展现了较为理想的结果。分析认为,PCA 在提取特征时将原始样本变量变为了彼此相互独立的主成分,原始变量相关程度越高,PCA 提取效果越好。各油样的光谱峰位置、峰强以及相对峰高基本一致(见图 1),经 Pearson 相关性分析发现,不同变量间相关系数在 0.9 以上。其高相关程度保证了 PCA 在提取油样光谱特征时的优越性。此外,相较于 MLP 模型,BDA 模型对各样本的识别与区分也相对较强。这与单一光谱模型、数据层融合模型的结果一致。

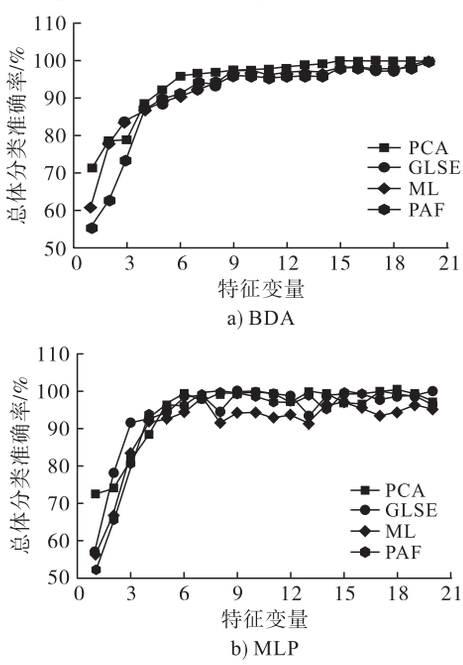


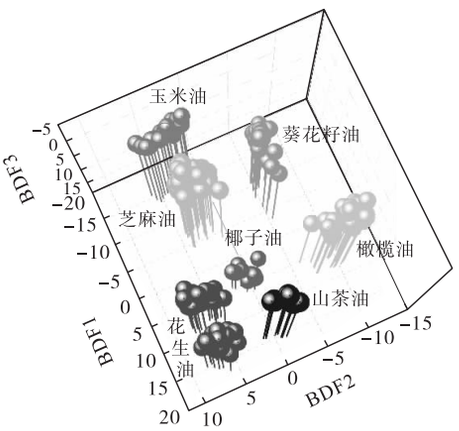
图 4　特征层融合模型对各样本的总体识别分类结果

2.5　植物油最佳识别模型

基于前述分析,最终选定基于 PCA 特征提取的油样特征层融合 BDA 模型为 180 份植物油样本的最佳识别模型。180 份植物油样本的最佳识别模型如图 5 所示。

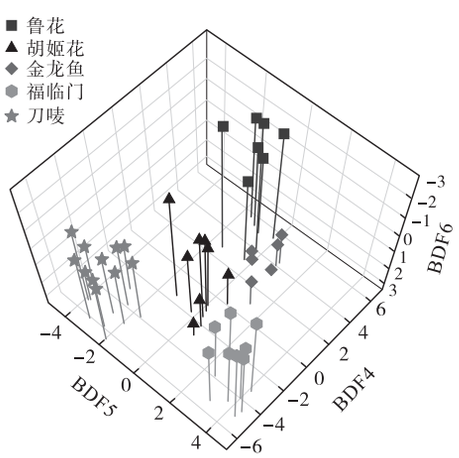
由图 5 可知,各样本之间彼此区分较为明显。在 BDF1 分类函数轴上,花生油与玉米油和芝麻油间区分明显,山茶油与椰子油区分明显,葵花籽油与

橄榄油区分明显;在 BDF2 分类函数轴上,花生油与葵花籽油、橄榄油、山茶油之间区分明显;在 BDF3 分类函数轴上,椰子油与其他样本区分明显。综合 3 个分类函数轴,各样本实现了 100% 的准确分类。相较于其他 6 种样本,花生油样本间的分布较为离散。鉴于此,通过最佳模型对花生油 5 种品牌的样本开展区分工作,得到了不同品牌花生油的分类结果,如图 6 所示。



注:BDF1、BDF2、BDF3 均为分类函数

图 5 180 份植物油样本的最佳识别模型



注:BDF4、BDF5、BDF6 均为分类函数

图 6 5 种品牌花生油的识别模型

由图 6 可知,5 种品牌的花生油较为清晰地区分开来。在 BDF4 分类函数轴上,鲁花与 3 种品牌样本(胡姬花、福临门、刀唛)彼此明显区分开来;在 BDF5 分类函数轴上,福临门与刀唛彼此区分较为明显;在 BDF6 分类函数轴上,鲁花与金龙鱼彼此明显区分。综合 3 个分类函数轴,各样本实现了 100% 的准确分类。结果表明采用基于 PCA 特征提取的特征层融合 BDA 模型可用于对同类样本品牌间的区分。

3 结 论

本实验采用衰减全反射 - 傅里叶红外光谱(ATR - FTIR)和表面增强拉曼光谱(SERS)分析技术获取市面常见的 7 种共计 180 份植物油样本的光谱信息数据,基于单一光谱模型、数据层融合模型、特征层融合模型,借助 BDA 和 MLP 两种方法构建模型,实现对各样本“种类 - 品牌”的两级识别分类工作。经比较最终选定基于 PCA 特征提取的特征层融合 BDA 模型为最佳识别模型,以此实现了 180 份植物油样本和 5 种品牌花生油的准确区分,结果较为理想。

在植物油样本的光谱数据模型中,基于融合方式构建的分类模型,对不同样本的区分能力是强于单一光谱数据集模型的,它是对多个信息源的结合,是对现有样本的“优化”。这种“优化”意味着模型拥有更为理想的识别精度和区分能力,这在法庭科学物证鉴定领域中是值得借鉴和应用的方向之一,对物证信息的全面刻画以及借助化学计量学构建更具客观性和描述性的模型有一定参考和借鉴意义。不同特征提取算法对模型的识别分类能力均有影响,本实验选择的 4 种常见算法中,PCA 在提取特征时,对相关程度较高的原始变量的提取效果较好。今后应进一步讨论和分析多种特征提取算法在原理和应用上的优劣,同时采集包含光谱技术在内的多种信息源(如色谱)的样本数据,为构建更为准确和可靠的数据分类模型提供一定基础,为实现对不同油样全面准确刻画分类提供一定参考。

参考文献:

[1] 景璐璐,马传国,闫亚鹏. 植物油中生物活性物质及其营养特性概述[J]. 中国油脂,2021,46(12):56 - 61.

[2] 刘凤霞,王莹,薛刚,等. 迷迭香脂溶性提取物在栀子油中的抗氧化性研究[J]. 中国油脂,2019,44(1):101 - 104.

[3] 杨丽萍,郭咪咪,段章群. 天然抗氧化剂迷迭香提取物在食用植物油中的应用研究进展[J]. 粮油食品科技,2022,30(2):95 - 100.

[4] 朱泉水,郝仕国,罗宁宁,等. 基于激光诱导荧光的植物油掺假检测与量化分析[J]. 中国激光,2019,46(12):273 - 278.

[5] 喻晴,钟培培,王远兴. 食用植物油掺假鉴别研究进展[J]. 中国油脂,2018,43(6):81 - 84,111.

[6] 时秋娜,刘占芳,朱军,等. 常见植物油的全二维气相色谱 - 质谱法检验[J]. 环境化学,2017,36(1):204 - 206.

[7] 郭莲仙,梁福睿,赵祖国,等. 基于稳定碳同位素技术的痕量动物油和植物油的区分检验研究[J]. 化学研究与应用,2014,26(8):1232 - 1236.

[8] HE X L, WANG J F, NIU F, et al. Characterization of heroin and its additives by attenuated total reflection (ATR) - Fourier transform infrared spectroscopy (FTIR) and multivariate analysis[J]. Anal Lett, 2020, 53(16): 2656 - 2670.

(下转第 89 页)