



中国油脂

China Oils and Fats

ISSN 1003-7969, CN 61-1099/TS

## 《中国油脂》网络首发论文

题目：基于机器学习算法的食用植物油掺伪鉴别的研究进展  
作者：孙婷婷，刘剑波，董界，周波，钟海雁  
网络首发日期：2020-09-23  
引用格式：孙婷婷，刘剑波，董界，周波，钟海雁. 基于机器学习算法的食用植物油掺伪鉴别的研究进展. 中国油脂.  
<https://kns.cnki.net/kcms/detail/61.1099.ts.20200923.0942.002.html>



**网络首发：**在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

**出版确认：**纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

# 基于机器学习算法的食用植物油掺伪鉴别的研究进展

孙婷婷<sup>1,2</sup> 刘剑波<sup>3</sup>, 董 界<sup>1,2</sup> 周 波<sup>1,2</sup> 钟海雁<sup>1,2</sup>

(1. 林产可食资源安全与加工利用湖南省重点实验室, 长沙 410004; 2. 中南林业科技大学 食品科学与工程学院, 长沙 410004; 3. 岳阳市质量计量检验检测中心食品检验所, 湖南 岳阳 414000)

**摘要:** 不同食用植物油的营养价值和功能价值不同, 导致市场上用低值低价油脂掺伪高值高价食用植物油的现象愈发普遍, 不仅损害了食用植物油脂生产者和消费者利益, 也不利于我国食用油脂产业的健康发展。由于掺伪油脂的种类多、掺伪手段复杂, 准确快速地对掺伪食用植物油进行鉴别具有一定的难度。近年来, 机器学习算法被应用到食用植物油掺伪鉴别的研究中, 取得了显著的研究成果。本文对现阶段国内外使用机器学习算法进行食用植物油掺伪鉴别的研究进展进行总结, 包括主成分分析、判别分析、支持向量机、随机森林、人工神经网络等算法, 并对各算法应用于食用植物油掺伪鉴别研究的优缺点进行了分析, 以期食用植物油掺伪鉴别的研究和应用提供一定的理论依据和方法参考。

**关键词:** 食用植物油; 掺伪鉴别; 机器学习; 算法

DOI:10.12166/j.zgyz.1003-7969/2020.00.100

## Research progress in adulteration identification of edible vegetable oils based on machine learning algorithms

SUN Tingting<sup>1,2</sup>, LIU Jianbo<sup>3</sup>, DONG Jie<sup>1,2</sup>, ZHOU Bo<sup>1,2</sup>, ZHONG Haiyan<sup>1,2</sup>

(1.Hunan Key Laboratory of Forestry Edible Sources Safety and Processing, Changsha 410004, China; 2.School of Food Science and Engineering, Central South University of Forestry and Technology, Changsha 410004, China; 3.Food Inspection Institute of Yueyang Quality Measurement Inspection and Testing Center, Yueyang Hunan 414000, China)

**Abstract:** Different edible vegetable oils have different nutritional and functional values, resulting in the gradually universal phenomenon using low-value and low-price oils to adulterate high-value and high-priced edible vegetable oils in the market, which not only damages the interests of edible vegetable oil producers and consumers, but also is harmful to the healthy development of the edible oil industry in China. Due to the variety of adulterated oils and the complicated adulteration methods, it is difficult to accurately and quickly identify adulterated edible vegetable oils. In recent years, machine learning algorithms have been applied to the adulteration identification of edible vegetable oils, and significant research results have been achieved. In this paper, the research progress of the application of machine learning algorithms to identify the edible vegetable oil adulteration is summarized, such as principal component analysis, discriminant analysis, support vector machine, random forest, and artificial neural network. The advantages and disadvantages of the machine learning algorithms in research on the adulteration identification of edible vegetable oils are also analyzed. This study can provide a theoretical basis and methodological reference for the research and application of adulteration identification of edible vegetable oils.

**作者简介:** 孙婷婷, 女, 1994 年出生, 硕士研究生, 食品营养与安全

**通信作者:** 钟海雁, 男, 1963 年出生, 教授, 森林食品加工与利用

周 波, 男, 1978 年出生, 副教授, 森林食品加工与利用

**Key words:** edible vegetable oil; adulteration identification; machine learning; algorithm

## 1 引言

我国常见的食用植物油包括花生油、大豆油、芝麻油、菜籽油、油茶籽油等。由于各类食用植物油营养价值和功能价值的不同,价格差距较大,导致用低值低价食用油掺伪高值高价食用油的现象屡见不鲜。目前食用植物油的掺伪方式主要有两种:一种是在高质高价食用植物油掺入低质低价的其他食用植物油;另一种是将有机溶剂浸出油掺到压榨油中,以次充好<sup>错误!未找到引用源。</sup>。国内外学者结合先进的实验仪器和检测技术,进行了食用油掺假鉴别的相关研究。基于先进的实验仪器和检测技术,研究人员会得到大量复杂的结构化量测数据。如何对这些数据进行分析,并从中提取出掺伪鉴别所需的信息,是决定掺伪检测鉴别结果的重要因素<sup>错误!未找到引用源。</sup>。同时,由于掺伪油脂的种类多、掺伪手段复杂,所以准确快速地对掺伪食用植物油进行鉴别具有一定的难度。

在食用植物油掺伪种类鉴别的研究中,指纹图谱相似度评价法的使用较为广泛,其是通过计算样本和纯植物油中特征成分含量的相似程度来实现的<sup>错误!未找到引用源。</sup>,但指纹图谱相似度评价法的局限性包括对指纹图谱中较小的共有峰识别能力较差、相似度的阈值难以客观确定等<sup>错误!未找到引用源。</sup>。在预测食用植物油掺伪量的研究中,线性回归、多项式回归、偏最小二乘回归等统计学回归分析方法使用较为广泛<sup>错误!未找到引用源。</sup>,但上述方法的局限性在于模型中所选的变量缺乏全面性以及复杂的高维非线性数据建模能力较差等<sup>错误!未找到引用源。</sup>。

机器学习是一门多领域综合交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多学科理论,其将计算机如何模拟及实现人类学习行为作为主要研究内容,探究计算机获取新知识、新技能的方式<sup>错误!未找到引用源。</sup>。在机器学习算法的指导下,计算机能够自动学习大量输入数据样本的数据结构和内在规律,有效地从复杂、高维、多变的大数据中挖掘出人类感兴趣的知识,从而对新样本进行智能识别,实现对未来的预测<sup>错误!未找到引用源。</sup>。将机器学习方法应用于掺伪鉴别研究中,能有效提取实验数据中隐藏的有价值的信息,利用训练好的模型自动对样本进行掺伪鉴别,可提高结果的准确性和客观性<sup>错误!未找到引用源。</sup>,所以基于机器学习算法的实验数据分析方法成为了食品科学领域的新兴研究热点,同时越来越多的机器学习算法被应用到食品掺伪鉴别研究中<sup>错误!未找到引用源。</sup>。但机器学习算法的种类繁多,各种算法在解决不同类型的数据分析问题上各有优势,所以在面对不同的掺伪检测需求时如何选择合适的机器学习算法,从而构建高质量的模型以提高植物油掺伪鉴别的质量,是目前食品分析领域内的研究热点。本文将对国内外基于机器算法的植物油掺伪鉴别的研究进展进行总结,对各种机器学习算法应用于植物油掺伪鉴别的优缺点进行分析,以期为后续研究提供参考和借鉴。

## 2 食用植物油掺伪鉴别中常用的机器学习算法

根据鉴别需求和问题性质,目前食用植物油掺伪鉴别的研究可大体分为三类,即鉴别样本是否掺伪、鉴别掺伪样本的掺伪种类、预测掺伪植物油的掺伪量。其中前两类对应于机器学习研究中的分类问题,最后一类对应于机器学习研究中的回归问题。

### 2.1 主成分分析

主成分分析法(Principal Component Analysis, PCA)是目前植物油掺伪分析研究中应用最为广泛的方法之一。目前应用 PCA 进行掺伪鉴别的基本研究思路是:通过对高维实验数

据进行降维处理,借助累计贡献率较高的两到三个主成分绘制样本的主成分得分图,以直观反应样本间的差异,进而推测食用植物油是否掺伪或掺伪油脂的具体种类<sup>错误!未找到引用源。</sup>。如冯利辉<sup>错误!未找到引用源。</sup>对 5 种食用植物油的光谱数据进行 PCA 分析,其二维主成分得分图表明同类样品基本聚集在相同区域,但对芝麻油和玉米油的分类效果较差。吴尧<sup>错误!未找到引用源。</sup>对 7 种食用植物油的酚类物质数据进行 PCA 分析,绘制了二维主成分得分图和热图,显示其能基本将各种植物油聚集在较小的区域内,但橄榄油、紫苏籽油、茶籽油、亚麻籽油、大豆油、菜籽油之间存在小部分重叠,分类效果欠佳。韩建勋等<sup>错误!未找到引用源。</sup>对 19 份油茶籽油、8 份大豆油、8 份菜籽油以及 8 份玉米油的光谱数据开展 PCA 分析,结果表明该方法很难区分上述四种食用油。

## 2.2 判别分析

植物油掺伪鉴别研究中使用的判别分析方法(Discrimination Analysis, DA)是一种有监督的分类算法,是化学计量建模中一种使用成熟的分析方法,主要包括线性判别分析算法(LDA)和偏最小二乘判别分析算法(PLS-DA)<sup>错误!未找到引用源。</sup>。目前应用 DA 进行掺伪鉴别的基本研究思路是:利用带标签的实验数据训练 DA 模型,利用方差贡献率较高的主因子构建判别函数得分图,计算样本的判别函数得分情况,通过设定得分阈值推测样本是否掺伪或掺伪油料的种类<sup>错误!未找到引用源。</sup>。如海铮等<sup>错误!未找到引用源。</sup>对掺入大豆油的油茶籽油和芝麻油样本的电子鼻检测数据开展了 LDA 分析,能够很好地鉴别掺伪芝麻油,但对掺伪油茶籽油的鉴别效果较差。钟培培<sup>错误!未找到引用源。</sup>对茶油、菜籽油、花生油样品的色谱数据进行了 PLS-DA 分析,对四种植物油样本的分类准确率达到 100%。Cheng 等<sup>错误!未找到引用源。</sup>通过二阶导数激光诱导荧光光谱法结合 DA 分析能够鉴别出三种纯油和两种混合油。Wang 等<sup>错误!未找到引用源。</sup>利用激发发射矩阵荧光光谱法结合二维 LDA 分析对掺假油茶籽油进行了分类,准确性较高。

## 2.3 支持向量机

支持向量机(Support Vector Machine, SVM)可用于解决分类问题<sup>错误!未找到引用源。</sup>。目前将 SVM 应用于植物油掺伪鉴别的相关研究较少,如章颖强<sup>错误!未找到引用源。</sup>利用掺入葵花籽油、大豆油、玉米油的橄榄油样品的拉曼光谱检测数据,结合最小二乘支持向量机进行了分类研究,可有效识别橄榄油中的掺伪油类型,对 36 个掺伪样本的掺伪种类的判别准确率达到 97%。毕艳兰等<sup>错误!未找到引用源。</sup>通过测定掺伪芝麻油样本的脂肪酸色谱数据,构建了基于支持向量机的芝麻油掺伪鉴别的两级分类模型,分别对样本是否掺伪和样本掺伪种类进行鉴别,并利用粒子群算法优化模型参数,两级分类模型的判别准确率分别达到了 100%和 98.7%。任小娜<sup>错误!未找到引用源。</sup>测得纯芝麻油和掺伪芝麻油的全样脂肪酸组成数据,构建了基于支持向量机的芝麻油掺伪判别模型,同样采用粒子群算法对模型参数进行优化,所得模型对芝麻油是否掺伪的鉴别准确率达到 100%。

## 2.4 随机森林

随机森林(Random Forest, RF)是一种由多棵树组成的集成分类器,在解决分类问题上相对于单个决策树具有很大的优势,目前已经有学者尝试使用随机森林算法进行植物油掺伪鉴别研究,如 Ai 等<sup>错误!未找到引用源。</sup>结合气质联用技术,对茶油、油菜籽油、玉米油、葵花籽油、芝麻油和橄榄油等六种植物油的脂肪酸采用 RF 法进行聚类分析,结果显示 RF 对样品种类的判别能力较好,且可以很好地定性和定量分析各脂肪酸变量对聚类模型的贡献;Xu 等<sup>错误!未找到引用源。</sup>采用多维气相色谱与飞行时间质谱结合技术,利用主成分分析和层次聚类分析以及 RF 算法建立了四种植物油的分类模型,可识别出掺假 5%大豆油的花生油并证明豆甾醇、 $\beta$ -香树脂醇、delta-7-豆甾烯醇、菜子甾醇、delta-7-菜油甾醇对判别结果影响大。



## 2.5 人工神经网络

人工神经网络 (Artificial Neural Network, ANN) 是由相互连接的大量神经元组成的复杂网络结构, 是对人脑中生物神经元的连接结构的抽象, 其训练过程模拟了人类大脑的生物神经网络对信息的处理、记忆的方式, 广泛应用于监督和非监督模式识别问题的解决<sup>错误!未找到引用源。</sup>。ANN 具有自学习、自组织、自适应以及很强的非线性函数逼近能力和强大的容错性, 是通用的非线性函数近似算法<sup>错误!未找到引用源。</sup>。目前已经有学者将 ANN 应用于解决食用植物油掺伪鉴别问题, 如荣茜等<sup>错误!未找到引用源。</sup>将茶油及其掺伪油样本的近红外光谱数据的主成分特征输入自组织映射神经网络, 构建了茶油掺伪鉴别的定性判别模型, 预测准确率达 95.33%, 优于马氏距离聚类分析法; Zhang 等<sup>错误!未找到引用源。</sup>使用 LDA 法对橄榄油的荧光光谱数据进行降维, 将降维后的数据输入 BP 神经网络, 构建了掺伪橄榄油的掺伪量定量预测模型, 其对掺花生油的橄榄油的掺伪量定量预测的均方根误差达到 1.34, 决定系数达到 0.996, 其预测效果优于 kNN 算法, 但与 PLSR 算法相比效果较差; Izquierdo 等<sup>错误!未找到引用源。</sup>利用橄榄油样本的热成像图训练卷积神经网络模型, 可对特级初榨橄榄油中掺入的精炼橄榄油、橄榄果榨油、葵花籽油进行识别, 算法判别准确率达到了 97% 以上; Rashvand 等<sup>错误!未找到引用源。</sup>将机器视觉技术与芝麻油掺伪检测结合起来, 利用人工神经网络模型芝麻油样本的颜色特征, 可实现对掺入不同植物油的芝麻油的掺伪量预测, 决定系数和均方误差分别达到了 0.944 和 0.0003; 方小伟<sup>错误!未找到引用源。</sup>利用 DAPCI-MS 与改进 BP 神经网络相结合的方法, 不仅能快速对油脂样品进行高通量分析, 还能快速准确地筛查出地沟油, 并对不同种类的食用油进行鉴定; Li 等<sup>错误!未找到引用源。</sup>基于径向基神经网络方法对油茶籽油的近红外光谱数据进行了建模, 对纯茶油是否掺伪鉴别的准确率高达 98.3%; Capar 等<sup>错误!未找到引用源。</sup>利用橄榄油样本的流变学参数数据来检测橄榄油掺假, 结果表明人工神经网络对橄榄油样本的掺伪种类的预测效果良好, 准确率达到 94.8%; 鲁小利等<sup>错误!未找到引用源。</sup>利用芝麻油和大豆油的电子鼻实验数据训练模糊神经网络模型, 可定量预测芝麻油和大豆油的混合物中芝麻油的含量, 绝对误差最大仅为 5.93%; 吴希军等<sup>错误!未找到引用源。</sup>利用掺伪芝麻油的三维荧光光谱数据, 结合广义回归神经网络 (GRNN) 对掺伪样本进行定量分析, 能够准确预测样本组分含量, 对掺伪样本掺伪量预测的相对误差小于 10%。

## 3 食用植物油掺伪鉴别中各机器学习算法优缺点分析

目前在食用植物油掺伪鉴别研究中, 主成分分析、判别分析的应用较为广泛, 支持向量机、随机森林、人工神经网络等算法尚未在该领域得到充分的应用和发展。虽然机器学习算法在一定程度上能够应用于鉴别和预测植物油是否掺伪、掺伪种类、掺伪量的问题, 但各算法在应用中有各自的优势和局限性, 实际研究中应当根据具体的问题需求和数据特性选取不同的算法。

### 3.1 主成分分析优缺点分析

利用主成分分析 (PCA) 进行植物油掺伪鉴别研究, 虽然操作简单并易于理解, 但鉴别效果欠佳。如利用 PCA 产生的得分图可有效检测数据集中的离群点<sup>错误!未找到引用源。</sup>, 但其只能对掺伪量较高的样本进行有效鉴别, 对低掺伪样本的鉴别能力较弱。在机器学习算法中, PCA 实际上并不专用于解决分类问题, 而是较多应用于数据降维任务和特征提取任务<sup>错误!未找到引用源。</sup>, 所以在掺伪鉴别研究中, 单独借助 PCA 进行植物油掺伪判别分析具有一定的局限性。基于此, 研究者常将 PCA 搭配其他算法共同使用, 首先利用 PCA 对数据进行降维, 在

低维空间中突出各类别样本的多元数据特性<sup>错误!未找到引用源。</sup>，再在此基础上进一步结合其他分类算法开展研究<sup>错误!未找到引用源。</sup>。同时，PCA 应用于植物油掺伪鉴别还有以下几点不足：针对掺伪鉴别数据中各变量间存在的复杂的非线性关系，作为线性降维算法的 PCA 很难做到高质量的分类<sup>错误!未找到引用源。</sup>；只利用方差贡献率排名靠前的两到三个主成分进行可视化分析，导致在累计贡献率不足时不能充分反映掺伪食用植物油的特征；PCA 是无监督的机器学习方法<sup>错误!未找到引用源。</sup>，当样本的组间差异不明晰而组内差异较大时，由于其对样本所属种类不加区分，难以发现样本的组间差异。

### 3.2 判别分析法优缺点分析

植物油掺伪鉴别研究中使用的判别分析算法（DA）属于有监督的机器学习分类算法<sup>错误!未找到引用源。</sup>，DA 在降维过程中考虑了样本的类别标签，突出了不同种类样本的差异<sup>错误!未找到引用源。</sup>，但对于低掺伪量的样本，由于其成分构成与纯油接近，在得分图上可能与纯油样本无法充分区别开，甚至存在重叠现象，所以 DA 方法对于低掺伪量的样本的分类效果不佳<sup>错误!未找到引用源。</sup>；同时，利用 DA 方法做出判定需要设置阈值，但阈值设定存在主观因素，且阈值高低直接影响了判别的准确性，因此这种方法相对缺乏客观性和科学性<sup>错误!未找到引用源。</sup>；此外，目前利用 DA 进行掺伪鉴别的研究多依托方差贡献值排名靠前的两到三个主因子来进行分析，若累计方差贡献率不足，会导致原数据部分信息损失，影响对样本分布描述的准确性。

### 3.3 支持向量机优缺点分析

支持向量机算法（SVM）利用少数支持向量确定的超平面进行分类，在面临非线性分类问题时，可通过引入核函数使 SVM 具备非线性分类能力<sup>错误!未找到引用源。</sup>。因此在多个基于 SVM 解决植物油掺伪鉴别问题的文献报道中，SVM 对非线性的掺伪鉴别问题的鉴别能力较好，对于低掺伪量样本同样具有一定的鉴别能力<sup>错误!未找到引用源。</sup>。但将 SVM 应用于掺伪鉴别也存在一定的局限性。首先，常规 SVM 只支持二分类问题，更适用于解决鉴别植物油是否掺伪的问题，对于鉴别植物油掺伪种类的多分类问题，则需要利用编码方法将问题进行转化，而编码方法的选择将在一定程度上影响分类的结果<sup>错误!未找到引用源。</sup>；其次，SVM 的模型训练操作复杂，参数选择过程既耗时又不易操作<sup>错误!未找到引用源。</sup>，且需要根据数据中样本和特征的具体情况为 SVM 选择合适的核函数，而核函数的选择是决定 SVM 性能的关键，对非线性分类问题没有通用的解决方案<sup>错误!未找到引用源。</sup>；第三，SVM 在模型训练时计算复杂度高，高昂的训练成本限制了 SVM 的广泛使用<sup>错误!未找到引用源。</sup>。

### 3.4 随机森林优缺点分析

随机森林算法（RF）属于集成学习算法（Ensemble Learning），运用了袋装技术（Bagging），通过结合多个树模型以降低泛化误差，相对于单个模型而言具有较强的鲁棒性和准确性<sup>错误!未找到引用源。</sup>。植物油掺伪鉴别研究中使用的数据集一般维度较高，而 RF 适合于处理高维数据集，且不必进行数据预处理<sup>错误!未找到引用源。</sup>；同时由于集成了大量的基学习器，RF 能够在一定程度上降低过拟合的风险性<sup>错误!未找到引用源。</sup>，并且 RF 能够根据训练好的模型来评估各个特征在分类过程中的重要性，在掺伪鉴别时可以筛选出植物油样本间存在显著差异的成分<sup>错误!未找到引用源。</sup>。现已经报道的相关文献表明，RF 在鉴别植物油掺伪的应用中准确率较高，并且在一定程度上克服了 SVM 的缺点<sup>错误!未找到引用源。</sup>。但 RF 的原理较为复杂，算法的空间和时间复杂度比较高<sup>错误!未找到引用源。</sup>，广泛应用于食品掺伪鉴别还有待进一步优化。

### 3.5 人工神经网络优缺点分析

人工神经网络（ANN）应用于食用植物油掺伪主要具有以下几点优势：ANN 解决高维非线性问题的能力较强<sup>错误!未找到引用源。</sup>，能准确描述繁多的植物油组分种类和组分之间存在的复

杂相关性，预测能力好，且对于低掺伪量的样本具有较强的分类能力<sup>错误!未找到引用源。</sup>；除给出样本的分类结果之外，ANN 还可给出样本分属各类别的概率，对低掺量样本的分类具有一定的参考价值。所以人工神经网络(ANN)应用于植物油掺伪检测的效果普遍较好<sup>错误!未找到引用源。</sup>。但 ANN 模型算法训练过程较为复杂，参数调整需要一定的经验，且网络层数、各层神经元数量等许多参数都会影响最终结果的质量<sup>错误!未找到引用源。</sup>；同时 ANN 模型的可解释性不强，对于植物油掺伪鉴别研究，ANN 只能给出分类结果，而难以解释是何种成分差异导致了鉴别结果<sup>错误!未找到引用源。</sup>；最主要的是 ANN 模型训练需要的数据量较大，在实验数据较少的情况下，ANN 模型的精度会受到影响<sup>错误!未找到引用源。</sup>。

表 1 植物油掺伪鉴别常用机器学习算法的优缺点分析

算法名称	优点	局限性	应用场景
主成分分析	应用广泛；原理简单；易于操作	鉴别准确率较低；非线性分类能力差；低掺样本鉴别能力差。	是否掺伪；掺伪种类
判别分析	原理简单；易于操作	鉴别准确率较低；分类存在主观性；低掺样本鉴别能力差。	是否掺伪；掺伪种类
支持向量机	可通过引入核函数作为非线性分类器；低掺样本鉴别能力较强；	问题编码方法影响分类结果；训练操作耗时且复杂；计算复杂度较高	是否掺伪；掺伪种类
随机森林	鉴别准确率较高；集成算法鲁棒性较强、不易过拟合；可适于高维非线性数据；可识别对掺伪鉴别有重要影响的特征；	原理复杂；计算复杂度较高	是否掺伪；掺伪种类
人工神经网络	鉴别准确率高；可适于高维非线性数据；低掺样本鉴别能力强；	训练操作耗时且复杂；计算复杂度较高；模型可解释性差；训练所需数据规模大	是否掺伪；掺伪种类；掺伪量预测

4 总结和展望

机器学习算法在食用植物油掺伪鉴别研究中已经有了一定的应用和成果，解决了掺伪鉴别研究中实验数据多维化、复杂化带来的分析难题，为掺伪鉴别研究开拓了新的方向。在后续植物油掺伪鉴别研究中，除了开发新的检测技术、研究新的特征指标物质外，如何合理利用机器学习算法对实验数据进行有效分析，提高掺伪鉴别的质量和效率，以服务于植物油掺伪鉴别应用，是食品分析领域一个重要的研究方向，具有极高的研究价值。基于本文的总结分析(见表 1)，为构建高质量的基于机器学习算法的掺伪鉴别模型，本文作者认为首先必须提供充足的高质量实验数据以保证模型训练的准确性；同时应结合实际情况，综合考量各因素的情况下选择合适的算法；充分运用交叉验证、外部验证等模型验证方法，在训练数据不足的情况下提高掺伪鉴别模型的准确率；尝试引入前沿算法以提高掺伪鉴别和预测模型的准确度，解决瓶颈问题。

参考文献：

[1] 赵淑娟, 郭平, 万建春, 等. 山茶油掺伪鉴别技术研究进展 [J]. 食品工业, 2018, 39(5):300-303.

[2] MUNCK L, NØRGAARD L, ENGELSEN S B, *et al.* Chemometrics in food science - a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance[J]. Chemometr Intell Lab, 1998, 44(1-2):31-60.

[3] 朱晓阳, 龙奇志, 钟海雁, 等. 基于脂肪酸和甘三酯组成的茶油掺伪检测参数的确定 [J]. 中国粮油学报, 2018, 33(4): 139-146.

- [4] 唐芳. 植物油脂脂肪酸指纹图谱的建立及掺假识别方法的研究 [D]. 湖南农业大学, 2011.
- [5] 孟庆华, 武真, 孙纪元, 等. 一种中药色谱指纹图谱相似度计算方法 [P]. 中国. CN 105388229 A. 2016-3-9.
- [6] 彭思敏. 基于脂肪酸组成变化的茶油掺伪判别及数学回归模型建立 [D]. 湖南农业大学, 2013.
- [7] 周波, 孙亚娟, 王进英, 等. 茶油挥发性成分指纹图谱的构建及掺伪定量检测方法研究[J]. 中南林业科技大学学报, 2017, 37(9):118-123.
- [8] 张东生, 金青哲, 王兴国, 等. 基于脂肪酸组成甄别油茶籽油掺伪的研究 [J]. 中国粮油学报, 2015, 30(1):124-128.
- [9] 王国丽, 陈晓飞, 刘刊, 等. 回归分析在水科学中的应用综述 [J]. 中国农村水利水电, 2004,11:40-44.
- [10] 欧华杰. 大数据背景下机器学习算法的综述 [J]. 中国信息化, 2019, (4):50-51.
- [11] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述 [J]. 模式识别与人工智能, 2014, 27(4), 327-336
- [12] 张润, 王永滨. 机器学习及其算法和发展研究 [J]. 中国传媒大学学报自然科学版, 2016, 23(2), 10-24
- [13] 王旻. 大数据背景下机器学习在数据挖掘中的应用 [J]. 信息与电脑(理论版), 2018, 415(21):143-144, 147.
- [14] 孙灵霞, 陈锦屏, 赵改名, 等. 化学计量学在食品分析中的应用研究进展 [J]. 食品工业科技, 2012, 33(7):407-411, 415.
- [15] BRO R, SMILDE A K. Principal component analysis [J]. Anal Methods-UK. 2014, 6(9): 2812–2831.
- [16] 冯利辉. 食用植物油掺伪检测与定量分析的近红外光谱法研究 [D]. 南昌大学, 2010.
- [17] 吴尧. 食用植物油多元掺伪鉴别技术研究 [D]. 中国农业科学院, 2016.
- [18] 韩建勋, 孙瑞雪, 陈颖, 等. FTIR 光谱结合化学计量学用于山茶油中掺杂大豆油的鉴别 [J]. 食品与发酵工程, 2019, 45(18):222-227.
- [19] BALLABIO D, CONSONNI V. Classification tools in chemistry. Part 1: linear models. PLS-DA [J]. Anal Methods-UK, 2013, 5(16):3790-3798.
- [20] 海铮, 王俊. 基于电子鼻山茶油芝麻油掺假的检测研究 [J]. 中国粮油学报, 2006, 21(3):204-209.
- [21] 钟培培. 色谱质谱联用技术结合化学计量学方法构建植物油品质鉴别体系 [D]. 南昌大学, 2018.
- [22] CHEN H, GENG D, CHEN T, *et al.* Second-derivative laser-induced fluorescence spectroscopy combined with chemometrics for authentication of the adulteration of camellia oil [J]. CYTA-J Food, 2018, 16(1):747-754.
- [23] WANG T, WU H, LONG W, *et al.* Rapid identification and quantification of cheaper vegetable oil adulteration in camellia oil by using excitation-emission matrix fluorescence spectroscopy combined with chemometrics [J]. Food Chem, 2019, 293:348-357.
- [24] HSU C, LIN C. A Comparison of Methods for Multiclass Support Vector Machines [J]. IEEE T Neural Networ, 2002, 13(2):415-425.
- [25] 章颖强. 食用植物油快速检测方法研究 [D]. 浙江大学, 2012.
- [26] 毕艳兰, 任小娜, 彭丹, 等. 粒子群最小二乘支持向量机结合偏最小二乘法用于芝麻油质量的鉴别 [J]. 分析化学, 2013, 41(9):1366-1372.
- [27] 任小娜. 芝麻油掺伪检测方法及体系模型的研究 [D]. 河南工业大学, 2012.
- [28] AI F, BIN J, ZHANG Z, *et al.* Application of random forests to select premium quality vegetable oils by their fatty acid composition [J]. Food Chem, 2014, 143:472-478.
- [29] XU B, ZHANG L, WANG H, *et al.* Characterization and authentication of four important edible oils using free phytosterol profiles established by GC-GC-TOF/MS [J]. Anal Methods-UK, 2014, 6(17):6860-6870.
- [30] BASHEER I A, HAJMEER M N. Artificial Neural Networks: Fundamentals, Computing, Design, and Application [J]. J Microbiol Meth, 2001, 43(1):3-31.



- [31] HAI Z, WANG J. Detection of adulteration in camellia seed oil and sesame oil using an electronic nose [J]. *Eur J Lipid Sci Tech*, 2006, 108(2):116-124.
- [32] 荣菡, 甘露菁, 王磊, 等. 基于近红外光谱的茶油掺伪快速检测方法的研究 [J]. *中国调味品*, 2019, 44(12):144-154.
- [33] ZHANG Y, LI T, CHENG H, *et al.* Excitation wavelength analysis of a laser-induced fluorescence technique for quantification of extra virgin olive oil adulteration. *Appl Optics*, 2019, 58(16):4484-4491.
- [34] IZQUIERDO M, LASTRA-MEJÍAS M, GONZÁLEZ-FLORES E, *et al.* Deep thermal imaging to compute the adulteration state of extra virgin olive oil [J]. *Comput Electron Agr.* 2020, 171.
- [35] RASHVAND M, AKBARNIA A. The feasibility of using image processing and artificial neural network for detecting the adulteration of sesame oil [J]. *AIMS Agr Food.* 2019,4(2):237-243.
- [36] 方小伟. 质谱法快速筛查劣质食用油的研究 [D]. 东华理工大学, 2016.
- [37] LI S, ZHU X, ZHAN J, *et al.* Authentication of Pure Camellia Oil by Using Near Infrared Spectroscopy and Pattern Recognition Techniques [J]. *J Food Sci*, 2012, 77(4):374-380.
- [38] CAPAR T D, KAVUNCUOGLU H, YALCIN H, *et al.* Rheological analysis for detection of extra virgin olive oil adulteration with vegetable oils: predicting oil type by artificial neural network [J]. *Qual Assur Saf Crop.* 2019, 11(8):687-699.
- [39] 鲁小利, 王俊. 仿生电子鼻在芝麻油掺伪检测中的应用研究 [J]. *粮食与油脂*, 2016, 29(6):75-77.
- [40] 吴希军, 崔耀耀, 潘钊, 等. 三维荧光光谱结合 Zernike 图像矩快速鉴别掺伪芝麻油 [J]. *光谱学与光谱分析*, 2018, 38(8):2456-2461.
- [41] 赵蕾. 主成分分析方法综述 [J]. *软件工程*, 2016, 19(6):1-3.
- [42] JOLLIFFE I T, CADIMA J. Principal component analysis: a review and recent developments [J]. *Phil Trans R Soc A*, 2016, 374(2065):20150202.
- [43] 王蕾, 李因果, 夏利宇. 基于正交迭代的有监督的稀疏主成分分析[J]. *统计与信息论坛*, 2018, 33(6):3-8.
- [44] QIU S, WANG J, GAO L. Discrimination and Characterization of Strawberry Juice Based on Electronic Nose and Tongue: Comparison of Different Juice Processing Approaches by LDA, PLSR, RF, and SVM [J]. *J Agr Food Chem*, 2014, 62(27):6426-6434.
- [45] SHI T, ZHU M, CHEN Y, *et al.* <sup>1</sup>H NMR combined with chemometrics for the rapid detection of adulteration in camellia oils [J]. *Food Chem*, 2018, 242:308-315.
- [46] BRERETON R G, LLOYD G R. Partial least squares discriminant analysis: taking the magic away [J]. *J Chemometr*, 2014, 28(4):213-225.
- [47] KUMAR B, VYAS O P, VYAS R. A comprehensive review on the variants of support vector machines [J]. *Mod Phys Lett B*, 2019, 33(25):1950303.
- [48] 章颖强, 董伟, 张冰, 等. 基于拉曼光谱和最小二乘支持向量机的橄榄油掺伪检测方法研究 [J]. *光谱学与光谱分析*, 2012, 32(6):116-120.
- [49] HUANG H, WANG Z, CHUNG W. Efficient parameter selection for support vector machines [J] *Enterp Inf Syst-UK*, 2019, 13(6):916-932.
- [50] AMARI S, WU S. Improving support vector machine classifiers by modifying kernel functions [J]. *Neural Networks*, 1999, 12(6):783-789.
- [51] SHI J, WEN Z, HE B, *et al.* Efficient Support Vector Machine Training Algorithm on GPUs [C]. 32nd AAAI Conference on Artificial Intelligence, 2018.
- [52] STROBL C, MALLEY J, TUTZ G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests [J]. *Psychol Methods*, 2009, 14(4):323-348.

