

植物油的红外光谱结合神经网络快速识别

接昭玮, 刘卓, 王继芬, 古昆山, 王之宇

(中国人民公安大学 侦查学院, 北京 100038)

摘要:为实现对植物油的快速检测,借助衰减全反射-傅里叶变换红外光谱分析技术并结合深度学习算法对植物油开展光谱模式识别工作。实验获取8种植物油样本的光谱数据,采用标准正态变换和一阶导数预处理方法消除背景干扰,同时采用竞争性自适应重加权算法模型对各样本特征光谱数据进行提取,分别建立长短记忆神经网络(LSTM)、基于Levenberg-Marquardt算法改进的BP神经网络对提取特征波长后的植物油种类进行预测识别与比较,并采用后者进行了实际样品的识别检测。结果表明,通过提取特征波长,可有效提高LSTM模型的识别准确率,其最优准确率从提取特征波长前的30%~40%提高到80%~90%,模型运行时间从提取特征波长前的111 min 25 s缩短至1 min 45 s。相较于LSTM模型,基于Levenberg-Marquardt算法改进的BP神经网络的分类识别准确率更高,达到99.852%,用于实际样品的识别,识别准确率达到100%。实验结果可为植物油的无损快速检验提供一定的参考与借鉴。

关键词:红外光谱;植物油;特征提取;机器学习;神经网络

中图分类号:TS225.1;O657.33 文献标识码:A 文章编号:1003-7969(2023)01-0079-06

Fast identification of vegetable oils by infrared spectroscopy combined with neural network

JIE Zhaowei, LIU Zhuo, WANG Jifen, GU Kunshan, WANG Zhiyu

(School of Investigation, People's Public Security University of China, Beijing 100038, China)

Abstract: In order to realize the fast identification of vegetable oil, attenuated total reflection-Fourier transform infrared spectroscopy technology and depth learning algorithm were used to carry out spectral pattern recognition of vegetable oil. The spectral data of 8 kinds of vegetable oil samples were obtained in the experiment. The standard normal variate and the first derivative preprocessing method were used to eliminate the background interference. At the same time, competitive adaptive reweighted sampling model was constructed to extract the features spectral data of each sample. Long short-term memory neural network (LSTM) and BP neural network improved by Levenberg-Marquardt algorithm were established to predict and compare the types of vegetable oil after extracting feature wavelength respectively, and the latter was used for identification of actual samples. The results showed that the recognition accuracy of the LSTM model could be effectively improved by extracting the feature wavelengths, and the best recognition rate increased from 30% - 40% before extracting the feature wavelengths to 80% - 90%, and the model running time was shortened from 111 min 25 s before extracting the feature wavelengths to 1 min 45 s. Compared with the LSTM model, BP neural network improved by Levenberg-Marquardt algorithm had a higher classification recognition accuracy of 99.852%, and the recognition accuracy was 100% when used for the recognition of actual samples. The

experimental results can provide some reference for nondestructive rapid testing of vegetable oils.

Key words: infrared spectroscopy; vegetable oil; feature extraction; machine learning; neural network

收稿日期:2021-12-07;修回日期:2022-10-11

作者简介:接昭玮(1998),男,硕士研究生,研究方向为刑事技术(E-mail)865571844@qq.com。

通信作者:刘卓,讲师,博士(E-mail)20052372@ppsuc.edu.cn。

食用植物油是人们日常生活中必不可少的烹饪用品,也是人体必需营养物质的主要来源之一^[1]。近年来,有些不法分子为牟取暴利对食用植物油的原料和制作工艺“投机取巧”,造成了重大食品安全隐患。

分子光谱分析技术是自 20 世纪 90 年代以来发展最快的分析技术,其以高效快速、制样简单以及无污染等独特的分析优点被广泛应用于食品品质分析中^[2-3]。当前,食用植物油的仪器检测方法主要有气相色谱-离子迁移谱、紫外光谱技术和 X 射线检测技术等,虽然这些方法测量精密度高,但存在仪器造价昂贵、处理步骤多、耗费时间长、检测效率低等缺陷,且试剂消耗大、易污染环境,不适合用于样品快速检测^[4]。红外光谱以其特征性好和适用样品范围广等特点,在法庭科学领域被广泛用于食品类物证的鉴别^[5-6]。

机器学习是一门关于数据学习的科学技术。当前,许多学者将传统机器学习模型应用于食用植物油中以开展掺伪鉴别研究工作^[7-9],但传统机器学习模型在面对大批量样本数据时,容易出现欠拟合现象,同时模型对异常数据的处理也十分困难。深度学习是建立于机器学习算法基础上的一种模式分析方法统称,神经网络作为深度学习应用的一个热门领域,具有较强的自学习功能和泛化能力,能够有效弥补传统机器学习模型在数据处理方面的不足。当前,深度学习在环境化学^[10]、材料分析^[11]、法庭科学^[12]等领域均有应用。将深度学习用于开展对光谱信息数据的挖掘,实现对样本检验结果的信息化和可视化是当下分析检测研究领域的热点之一。

本文提出通过红外光谱结合深度学习算法对食用植物油进行分类识别处理。采用衰减全反射-傅里叶变换红外光谱(ATR-FTIR)分析技术,获取 8 种常见的食用植物油光谱数据,通过竞争性自适应重加权算法(CARS, competitive adaptive reweighted sampling)提取各样本的特征光谱数据,以缩短检测时间,提高分类准确率,在此基础上构建长短记忆神经网络、基于 Levenberg-Marquardt 算法改进的向后传播算法(BP)神经网络开展对不同植物油的快速无损分类识别工作,并将后者用于实际油样的识别,测定识别准确率,以期为食用植物油的快速无损检测提供参考。

1 材料与方法

1.1 实验材料

从市场上收集 8 种共计 160 个常见的食用植物油样本(见表 1)。

表 1 食用植物油样本

类别	样本数量
小磨香油	20
花生油	20
葵花籽油	20
山茶油	20
橄榄油	20
亚麻籽油	20
花椒籽油	20
椰子油	20

Nicolet is10 型傅里叶变换红外光谱仪、衰减全反射附件(金刚石晶体),美国 Thermo Fisher Scientific 公司。

1.2 实验方法

1.2.1 光谱数据采集

利用衰减全反射-傅里叶变换红外光谱(ATR-FTIR)分析技术采集植物油的光谱数据。采集条件:光谱采集范围 $4\ 000 \sim 650\ \text{cm}^{-1}$,光谱分辨率 $2\ \text{cm}^{-1}$,扫描次数 64 次。

1.2.2 光谱数据预处理

为确保实验数据分析的准确性,排除仪器放置环境以及机器本身产生的噪声干扰,对所测得的红外光谱数据进行平滑降噪处理;为消除固体颗粒大小、表面散射以及光程变化对漫反射光谱影响,对测得的光谱进行标准正态变换(SNV, Standard normal variate)处理;为有效消除基线和其他背景等干扰,使重叠峰发生分离,再对经平滑降噪、SNV 处理后的光谱进行一阶导数(DT)处理。

1.2.3 特征光谱数据的提取

CARS 是基于自适应重加权采样(ARS)技术,选择偏最小二乘法(PLS)模型中回归系数绝对值大的波长点,利用交互验证选出均方标准差(RMSECV)值最低的子集,去掉权重小的波长点,最终有效寻出最优变量组合^[13]。

实验在文献[14-15]基础上建立 CARS 模型,采用十折交叉验证,将采样的总次数设置为 50 次,以得到相关样本光谱变量筛选图,从而得到特征光谱。

1.2.4 长短记忆神经网络(LSTM)

LSTM 是在循环神经网络(RNN)基础上研发的一种改进式循环神经网络,它不仅可以有效地处理序列数据,还可以选择性地存储信息,从而弥补传统 RNN 在解决长期依赖问题中的局限性。参考文献[16-17]使用 MATLAB 搭建 LSTM 模型,分别将 8 种食用植物油特征波长提取前后的光谱数据导入

LSTM 模型进行训练验证,以对 8 种食用植物油样本进行分类识别研究,同时采用 CARS 对提取特征波长前后分类结果提升的有效性和必要性进行验证。

1.2.5 基于 Levenberg - Marquardt (LM) 算法改进 BP 神经网络

参考文献[18 - 19]建立基于 LM 算法改进 BP 神经网络模型进行食用植物油的分类识别。

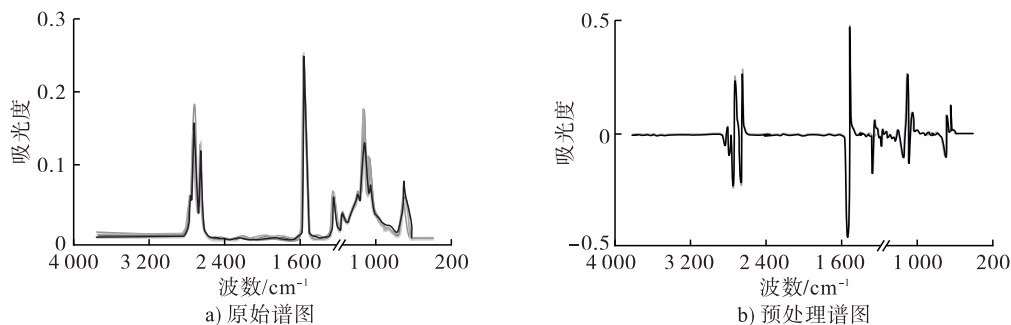


图1 食用植物油预处理前后红外光谱图的比较

2.2 各样本光谱数据特征波长提取

特征波长优选过程以及食用植物油光谱变量筛选图如图 2 所示。由图 2a 可知,特征波长数量随着样本运行次数的增加而减少;图 2b 表示十折交叉验证均方标准差 (RMSECV) 值的变化趋势;图 2c 表示

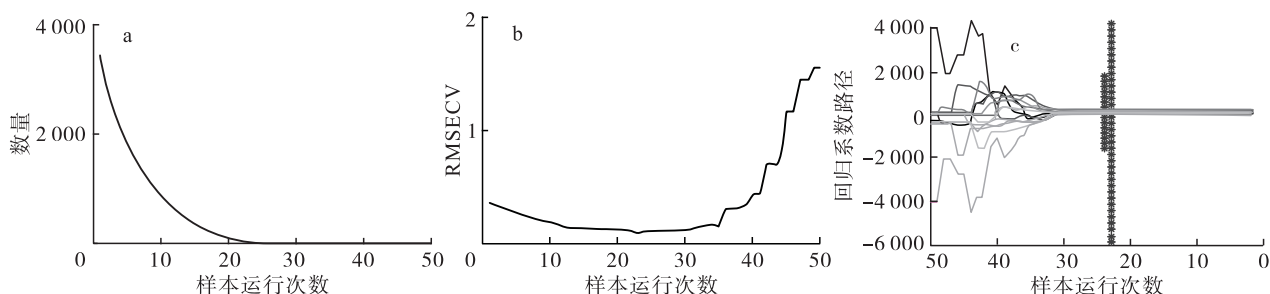


图2 特征波长优选过程及食用植物油光谱变量筛选图

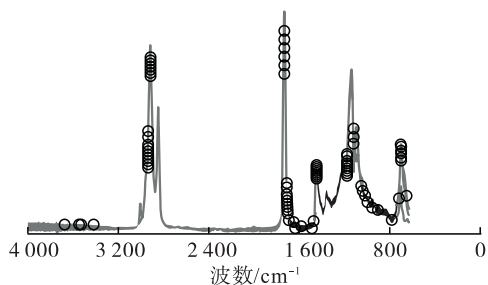


图3 光谱特征波长提取结果

2.3 长短记忆神经网络

LSTM 对于未提取特征波长的 8 种食用植物油分类识别结果如图 4 所示,对于提取特征波长的 8 种食用植物油分类识别结果如图 5 所示。

由图 4 可知,在未提取特征波长时,LSTM 模型

2 结果与分析

2.1 光谱数据预处理

将经过平滑降噪、SNV、DT 预处理后的食用植物油光谱结果与原始谱图对比,结果如图 1 所示。由图 1 可知,预处理后的光谱能够消除由固体颗粒大小、表面散射以及光程变化对漫反射光谱的影响,有效提高了谱图的分辨率和灵敏度。

每个变量回归系数路径的变化。将光谱数据连续运行 23 次,取 23 次结果中频率最高的波长作为真实特征波长。基于 CARS 模型提取,实验共提取出 8 种植物油的特征波长 105 个,光谱特征提取波长结果如图 3 所示。

对于实验中食用植物油的分类识别准确率基本维持在 10% 以上,最优准确率在 30% ~ 40% 的区间内,模型运算时间长达 111 min 25 s。由图 5 可知,提取特征波长后 LSTM 模型对于各样本的分类识别准确率在仪器稳定后基本维持在 60% 以上,最优准确率提升至 80% ~ 90% 的区间内,模型运算时间仅为 1 min 45 s。表明通过 CARS 对 8 种食用植物油进行特征波长提取可以有效提高模型分类识别准确率以及缩短模型的运行时间,从而降低模型的计算复杂度。

但从分类结果看,LSTM 模型的最优识别结果准确率仍然相对较低,结果稳定性方面仍具有提升空间。

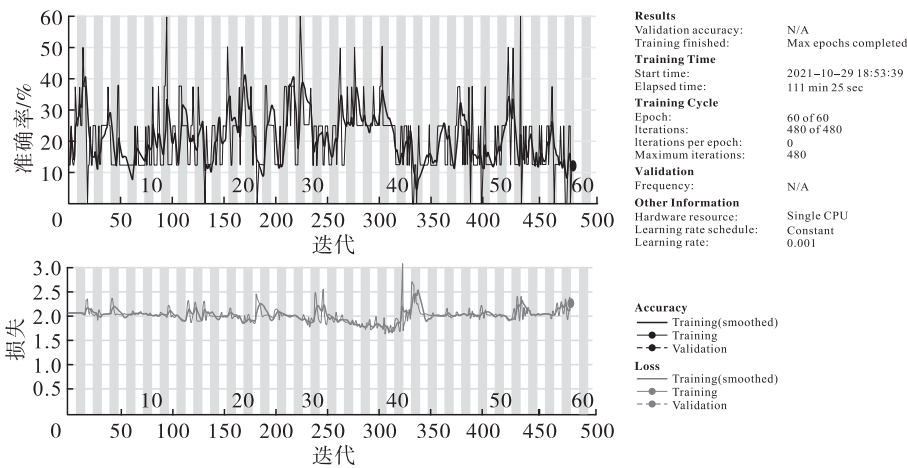


图4 LSTM对于未提取特征波长的8种食用植物油分类识别结果

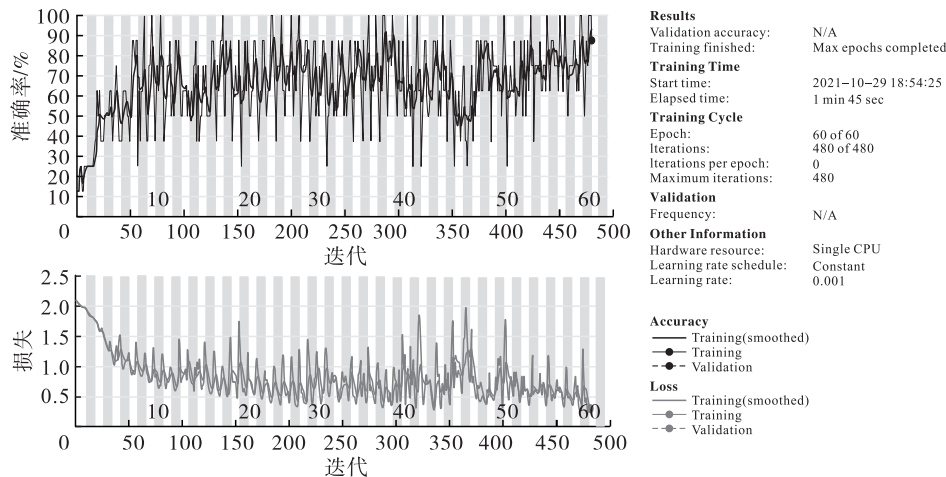


图5 LSTM对于提取特征波长的8种食用植物油分类识别结果

2.4 基于LM算法改进BP神经网络

前馈神经网络中的向后传播算法(BP)具有逼近任意非线性网络连续映射的能力,在机器学习领域内有着广泛应用^[20]。但是BP神经网络具有自身的局限性,主要表现在收敛速度慢、数值稳定性差以及参数难以调整等方面^[21]。LM算法是最优化算法中的一种,是以最广泛的非线性最小二乘算法为基础,利用梯度求最大(小)值的方法。LM算法的优势在于对过参数化问题不敏感,能够有效处理参数冗余问题,使损失函数进入局部极小值的可能性降低。利用LM算法改进的BP神经网络,可有效克服传统BP神经网络中的局限性,从而提高机器学习分类识别的准确率。

将样本数据按照训练集和测试集7:3的比例导入建立的模型中进行训练,训练过程如图6所示。由图6可看出,训练集在训练190次后达到收敛,测试集在训练220次后达到收敛,此时会取得模型最优的识别效果。

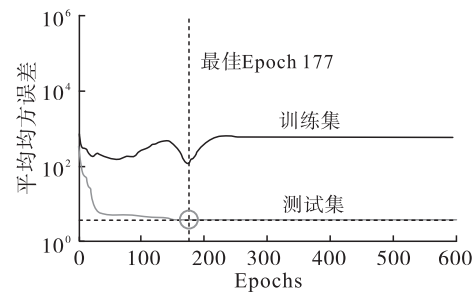


图6 样本在基于LM算法改进BP神经网络中的训练曲线

模型中训练集和测试集数据的拟合情况如图7所示,其拟合过程反映了机器学习的预测值与实际值的相似程度,拟合度越高说明机器学习的预测识别效果越好。由图7可知,训练集数据与原始数据的拟合程度达到99.921%,测试集数据与原始数据的拟合程度达到99.859%,模型总体拟合程度达到99.852%。相较于LSTM模型,基于LM算法改进BP神经网络对于食用植物油种类间的识别率较高。

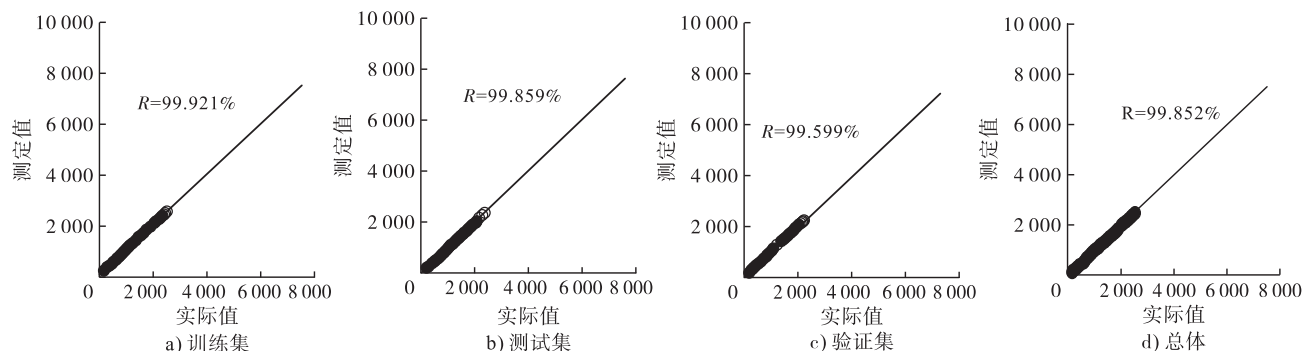


图 7 基于 LM 算法改进 BP 神经网络的最终结果

2.5 实际样品的分类识别

以 8 种植物油(小磨香油、花生油、葵花籽油、山茶油、橄榄油、亚麻籽油、花椒籽油和椰子油)共 160 份实验样本作为训练集,以 24 份已知油样作为测试集,在测得已知油样的傅里叶变换红外光谱图后,采用 CARS 对其进行特征波长提取,并以建立的基于 LM 算法改进 BP 神经网络对该 24 份已知油样开展分类识别工作,结果如图 8 所示。

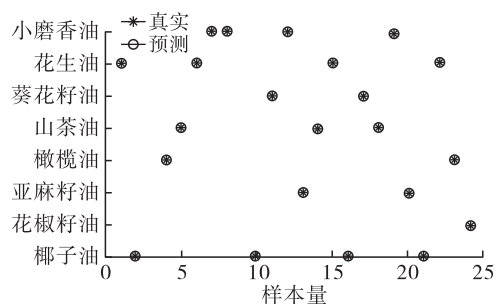


图 8 基于 LM 算法改进 BP 神经网络对植物油种类预测识别结果

由图 8 可知,基于 LM 算法改进 BP 神经网络对各样本预测识别的准确率达到 100%。

3 结论

实验采用衰减全反射-傅里叶变换红外光谱分析技术获取各类植物油样本的光谱数据信息,首先利用 CARS 对各样本的光谱数据特征波长进行提取,建立 LSTM 模型对特征波长提取前后的样本数据开展分类识别工作,并针对 LSTM 模型的局限性,建立基于 Levenberg-Marquardt 算法改进的 BP 神经网络模型对提取特征波长后的植物油进行分类,以获得最优的光谱化学识别效果。结果表明,基于 CARS 特征波长提取构建的 LSTM 模型识别最优准确率达到 80%~90%,同时相较于原始数据的分类结果,其运算时间缩短至 1 min 45 s,表明开展特征波长的提取工作对于食用植物油的光谱进行模式识别是十分有必要的,同时基于 LM 算法改进的 BP 神经网络使食用植物油样本的总体分类精度达到了 99.

852%,能够有效克服 LSTM 在数据并行处理方面存在的不足。这对于今后的食品安全监测工作来说是值得考虑的方向之一,实验为今后深度学习在食品安全领域中的应用提供了方向与指导。

参考文献:

- [1] 薛莉,黄晓荣,汪雪芳,等.食用植物油营养成分及检测技术的研究进展[J].食品安全质量检测学报,2017,8(2):446-451.
- [2] 江秀明,周长智,李建伟,等.可见分光光度法测定食用油过氧化值[J].郑州粮食学院学报,1999(2):57-59.
- [3] MUIK B, LENDL B, MOLINA-DIAZ A, et al. Direct monitoring of lipid oxidation in edible oils by Fourier transform Raman spectroscopy[J]. Chem Phys Lipids, 2005, 134(2):173-182.
- [4] 姚晶,樊婷,苏春燕,等.食用植物油组分分析快速检测技术研究进展[J].粮油食品科技,2020,28(2):97-102.
- [5] BAIANO A. Applications of hyperspectral imaging for quality assessment of liquid based and semi-liquid food products: a review[J]. J Food Eng,2017,214:10-15.
- [6] 刘亚超,李永玉,彭彦昆,等.近红外二维相关光谱的掺和大米判别[J].光谱学与光谱分析,2020,40(5):1559-1564.
- [7] 俞玥,宋嘉慧,鲁玉杰,等.食用植物油掺伪鉴别技术研究进展[J].食品安全质量检测学报,2021,12(13):5153-5161.
- [8] 孙婷婷.基于机器学习算法的油茶籽油掺伪鉴别模型的建立[D].长沙:中南林业科技大学,2021.
- [9] 王满琴.基于反射率光谱和深度学习网络的食用油饱和脂肪酸检测[D].合肥:安徽大学,2021.
- [10] 但德忠,沈璐,祝艳涛.环境样品分析[J].分析试验室,2006(6):96-122.
- [11] 万群.新材料的发展与分析化学[J].分析试验室,1993(1):30-38.
- [12] HOFFMAN B L, HACKMAN L, LINDENFELD L A. Training for communication in forensic science[J]. Emerg Topics Life Sci, 2021, 5(3): 359-365.

(下转第 93 页)

- [10] 王蕤, 汤富彬, 钟冬莲, 等. 4种胡桃科坚果中氨基酸和脂肪酸组成分析与营养评价[J]. 中国油脂, 2020, 45(4): 86-91.
- [11] URSIN V M. Modification of plant lipids for human health: development of functional land-based *omega*-3 fatty acids[J]. J Nutr, 2003, 133(12): 4271-4274.
- [12] LIU H L, YIN Z J, XIAO L, et al. Identification and evaluation of ω -3 fatty acid desaturase genes for hyperfortifying α -linolenic acid in transgenic rice seed[J]. J Exp Bot, 2012, 63(8): 3279-3287.
- [13] CHEN W, GONG L, GUO Z L, et al. A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics[J]. Mol Plant, 2013, 6(6): 1769-1780.
- [14] 钱瑞, 任锡毅, 刘永翔, 等. 基于广泛靶向代谢组学的竹黄活性成分分析[J]. 菌物学报, 2019, 38(5): 631-645.
- [15] XU L, XU Z Z, WANG X, et al. The application of pseudotargeted metabolomics method for fruit juices discrimination[J/OL]. Food Chem, 2020, 316: 126278 [2021-09-17]. <https://doi.org/10.1016/j.foodchem.2020.126278>.
- [16] FRAGA C G, CLOWERS B H, MOORE R J, et al. Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography-mass spectrometry, XCMS, and chemometrics[J]. Anal Chem, 2010, 82(10): 4165-4173.
- [17] 王林. 一种高效生产L-脯氨酸的方法: CN201910414709.9 [P]. 2019-08-09.
- [18] 缪正兴, 张仲明, 李宝忠. L-脯氨酸的生产及其应用[J]. 发酵科技通讯, 2004, 33(2): 21-22.
- [19] CHACON J V, ARBELAEZ M I A, JORGE J H, et al. pH-responsive poly(aspartic acid) hydrogel-coated magnetite nanoparticles for biomedical applications[J]. Mat Sci Eng C, 2017, 77(3): 366-373.
- [20] MATHUR N, GOSWAMI G K, PATHAK A N. Structural comparison, docking and substrate interaction study of modeled *endo*-1, 4-*beta* xylanase enzyme of *Bacillus brevis* [J]. J Mol Graph Model, 2017, 74(2): 337-343.
- [21] 胡艳丽, 马拥军, 罗庆平, 等. L-天门冬氨酸诱导叠层碳酸钙微晶的形成[J]. 人工晶体学报, 2010, 39(3): 802-806, 812.
- [22] 邵江娟, 王唯, 陈斌寅, 等. 以牡蛎壳为钙源的L-天冬氨酸螯合钙的合成及其生物利用度研究[J]. 南京中医药大学学报, 2012, 28(4): 371-373.
- [23] 董学伟, 张春枝, 金凤燮. 中空纤维膜反应器转化L-天冬氨酸生成L-丙氨酸[J]. 食品与发酵工业, 2003, 29(1): 32-35.
- [24] GARGIULO A, ACUNA A, GARGIULO M, et al. Effects of cycloleucine in the nucleus accumbens septum the elevated plus maze test in rats[J]. Neuropsychobiology, 2020, 79(3): 191-197.
- [25] 张春娥, 张惠, 刘楚怡, 等. 亚油酸的研究进展[J]. 粮油加工, 2010(5): 18-21.
- [26] 范亚苇, 邓泽元, 刘蓉, 等. 共轭亚油酸对动脉粥样硬化老龄大鼠血脂和血浆脂肪酸的影响[J]. 营养学报, 2006(6): 472-474, 478.
- [27] 方志玲, 沈继红, 刘发义, 等. 共轭亚油酸对小鼠免疫功能的影响[J]. 中国海洋药物, 2003(5): 23-26.
- [28] 衣丹, 林学政, 沈继红, 等. 共轭亚油酸降血脂及抗动脉粥样硬化作用的研究[J]. 现代生物医学进展, 2011, 11(7): 1228-1230.
- [29] 吴端, 王力军, 杨仕梅, 等. 植物种子 α -亚麻酸形成及调控机理研究进展[J]. 植物遗传资源学报, 2020, 21(1): 53-66.
- [30] KÜLLENBERG D, TAYLOR L A, SCHNEIDER M, et al. Health effects of dietary phospholipids [J]. Lipids Health Dis, 2012, 11: 3-18.
- [31] 安红, 宋伟明, 张宏波. 磷脂化学及应用技术[M]. 北京: 中国计量出版社, 2006: 85-170.
- (上接第83页)
- [13] 罗一甲, 祝赫, 李潇涵, 等. 赤霞珠酿酒葡萄总酚含量的近红外光谱定量分析[J]. 光谱学与光谱分析, 2021, 41(7): 2036-2042.
- [14] 吴剑飞. 基于CARS-PLSR的冬小麦叶片含水量快速无损检测研究[J]. 赤峰学院学报(自然科学版), 2021, 37(7): 22-26.
- [15] 王海龙, 杨国国, 张瑜, 等. 竞争性自适应重加权算法和相关系数法提取特征波长检测番茄叶片真菌病害[J]. 光谱学与光谱分析, 2017, 37(7): 2115-2119.
- [16] 史运涛, 任鹏, 李书钦, 等. 基于长短期记忆神经网络的重大活动中鲜(冻)肉制品铅含量风险预测[J]. 食品安全质量检测学报, 2022, 13(7): 2326-2333.
- [17] 康雁, 徐玉龙, 寇勇奇, 等. 基于Transformer和LSTM的药物相互作用预测[J]. 计算机科学, 2022, 49(S1): 17-21, 205.
- [18] 邹艳均, 武周虎, 任鹏, 等. 人工蜂群优化LM-BP网络在东平湖水水质评价中的应用[J]. 水电能源科学, 2022, 40(4): 66-69, 36.
- [19] 简晓春, 王利伟, 闵峰. 基于LM算法的BP神经网络对汽车排放污染物的预测[J]. 重庆理工大学学报(自然科学版), 2012, 26(7): 11-16.
- [20] 郑毅. 时间序列数据的胶囊式LSTM特征提取算法研究[D]. 上海: 华中师范大学, 2018.
- [21] 柳益君, 古春生, 朱广萍, 等. LM神经网络在环境质量评估中的应用[J]. 江南大学学报(自然科学版), 2010, 9(2): 213-216.