

基于 KPCA 与 KLPP 及 Wilks 统计量的留兰香 三维荧光数据特征提取与鉴别分析

殷 勇，徐非凡，于慧春，袁云霞

(河南科技大学食品与生物工程学院，洛阳 471000)

摘 要：为实现留兰香产地的快速鉴别，该研究提出了一种核主成分分析（kernel principal component analysis, KPCA）与核局部保持投影（kernel locality preserving projections, KLPP）及 Wilks Λ 统计量序贯融合的特征波长提取策略，在此基础上鉴别 5 个产地的留兰香。首先，在采集 5 个产地 300 个留兰香样本的三维荧光数据后，运用三角形内插法去除原始光谱中的瑞利散射和拉曼散射，并运用 SG（Savitzky-Golay）对数据进行平滑预处理。然后，对预处理后的荧光光谱数据分别利用 KPCA、KPCA+KLPP、KPCA+Wilks Λ 统计量、KPCA+KLPP+Wilks Λ 统计量 4 种方法提取特征激发波长和特征发射波长。接着，按特征激发波长从小到大顺序将其对应的特征发射波长光谱值首尾相连转换成行向量；4 种方法从 300 个样本中各得到 1 个 300 行的特征波长光谱值矩阵。再者，运用 Fisher 判别分析（fisher discriminant analysis, FDA）对特征波长光谱值矩阵进行数据可分性融合，生成可分性 FD（fisher discriminant）变量。选取前 4 个累计判别能力达到 99% 的 FD 变量作为鉴别模型的输入向量。最后，用支持向量机（support vector machine, SVM）算法分析 4 个 FD 变量，分别得到对应于 4 种特征提取波长方法的 FDA+SVM 鉴别结果，其正确率分别为 92.00%、96.00%、94.67%、100%。结果表明，所提出的 KPCA+KLPP+Wilks Λ 统计量序贯融合的特征波长提取策略能够有效减少三维荧光光谱数据的冗余，并能表征原始荧光数据的信息特征，实现了 5 种留兰香产地的正确鉴别。该研究可为后续利用三维荧光光谱开展留兰香重要组分量化分析提供一定的基础。

关键词：荧光光谱；判别分析；模型；留兰香；核主成分分析

doi: 10.11975/j.issn.1002-6819.202405074

中图分类号: TS207.3

文献标志码: A

文章编号: 1002-6819(2024)-19-0272-09

殷勇, 徐非凡, 于慧春, 等. 基于 KPCA 与 KLPP 及 Wilks 统计量的留兰香三维荧光数据特征提取与鉴别分析[J]. 农业工程学报, 2024, 40(19): 272-280. doi: 10.11975/j.issn.1002-6819.202405074 <http://www.tcsae.org>

YIN Yong, XU Feifan, YU Huichun, et al. Spearmint three-dimensional fluorescence data feature extraction and identification based on KPCA and KLPP and Wilks statistics [J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(19): 272-280. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202405074 <http://www.tcsae.org>

0 引 言

留兰香是一种药食同源植物，在中国的许多地区都有种植，可以用来治疗感冒、头痛、发热等疾病，留兰香的嫩枝、叶多被作为调味香料^[1]。留兰香的挥发油含量较高，广泛应用于食品、日化等领域^[2-4]。基于此，人们对留兰香的需求正在不断增加。但市场上的留兰香品质参差不齐，以次充好现象时有发生，严重影响了留兰香市场正常运行。因此，为了鉴别真伪，控制品质，监管市场，需要研究一种高效的留兰香鉴别方法。

目前市场上尚无一种快速有效的鉴别方法鉴别不同产地的留兰香。在食品与中药领域中，光谱法^[5-7]、质谱法^[8-10]、色谱法等^[11-13]是较为常见的检测方法，其中三维荧光光谱技术由于其简单、高效等特点而备受青睐。相比于常规荧光图谱，三维荧光图谱能够反映相应激发

波长及发射波长下的荧光强度变化，可以将荧光物质的荧光特性更加完整地描述出来^[14-15]。例如，张轩等^[16]通过对宁夏枸杞果实的三维荧光光谱特性研究实现了枸杞子的产地鉴别；鹿鸣^[17]在不同产地稻种的荧光光谱研究中分析了其中重要化学成分的光谱特征如核黄素、叶绿素等。而香芹酮、黄酮、柠檬烯和维生素 C 等是留兰香产生荧光的主要成分，三维荧光图谱能够清晰地展现出它们的荧光峰位置、荧光强度及其变化趋势，因此可以通过分析三维荧光图谱来进行留兰香的产地鉴别。

特征提取是三维荧光光谱数据分析的关键^[18]，主要分为参量化方法和数据降维方法。前者通过提取光谱的特征参数对原始光谱数据进行表征，并利用这些参数进行后续的分类识别，但这种方法忽略了不同特征之间存在的内在联系^[19]。数据降维方法则是通过映射获得高维数据的低维表示，这种方法可以有效避免维数灾难和减少高维空间中的冗余信息，进而促进高维数据的分类、压缩及可视化^[20-21]。例如，李小占等^[22]通过数据降维方法实现了哈密瓜表面缺陷的无损检测；周博等^[23]通过 PCA（principal component analysis）算法进行光谱特征

收稿日期: 2024-05-13 修订日期: 2024-08-03

基金项目: 国家重点研发计划项目 (2017YFC1600802)

作者简介: 殷勇, 教授, 博士生导师, 研究方向为农产品、食品品质检测。

Email: yinyong@haust.edu.cn

提取和降维,实现了不同烟草组分的分类判别。但 PCA 在进行多组分鉴别时效果并不理想,因此,基于数据降维方法提出一种 KPCA (kernel principal component analysis) 与 KLPP (kernel locality preserving projections) 及 Wilks Λ 统计量序贯融合的特征波长提取策略^[24-25],该方法可以有效地降低荧光数据的冗余性,提高留兰香的鉴别准确率。

本文采用 KPCA 与 KLPP 及 Wilks Λ 统计量序贯融合方法作为留兰香三维荧光数据特征波长提取策略;用 FDA (fisher discriminant analysis) 算法构建鉴别模型输入向量;最后运用 SVM (support vector machine) 算法对留兰香产地进行分类识别。该三维荧光数据特征波长提取策略不仅能够有效减少光谱数据的冗余,还能实现留兰香产地的正确鉴别。

1 材料与方法

1.1 材 料

选取河南、安徽、北京、上海和广东 5 个产地的留兰香进行试验。每个产地的留兰香采集 60 个样本,从中随机选择 45 个组成训练集,剩下的 15 个组成测试集。训练集共计 225 个样本,测试集共计 75 个样本。

1.2 仪 器

荧光数据的采集用实验室自制的荧光光谱仪,具体结构及主要性能参数见文献^[26],主要部件及其参数如表 1 所示。

表 1 荧光光谱仪主要部件及参数

Table 1 Main components and parameters of fluorescence spectrometer

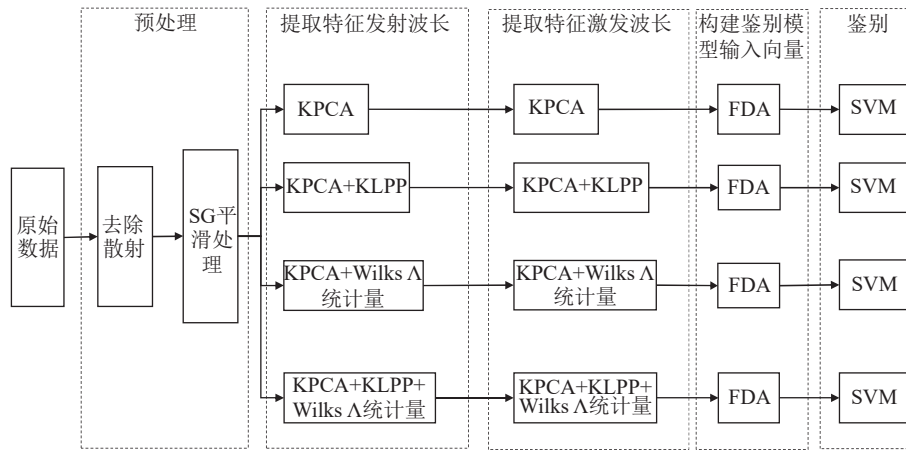
部件 Component	型号 Model	规格 Specification
紫外增强型氙灯 Ultraviolet-enhanced xenon lamp	LSB-X150AUV	150 W, 7.5 A, 20 V, 波长准确度: ± 0.5 nm
激发单谱仪 Excitation single spectrometer	SmartFluo-Pro-DZ-J	激发范围: 200~800 nm, 带宽: 5 nm, 波长重复性指 数: ± 1 nm
光栅 Raster	SmartFluo-Pro-DZ-L	1 800 g $\cdot$ mm $^{-1}$, 400 nm, 体相全息
CCD 探测器 CCD detector	SmartFluo-Pro-DZ-P	发射范围: 200~800 nm, 带宽: 5 nm

1.3 荧光数据采集

荧光光谱仪开机后预热 30 min。将留兰香样本展平后放入比色皿中进行荧光数据的采集^[27]。设置激发波长范围 350~750 nm, 步长 5 nm, 发射波长范围 500~800 nm, 步长 0.4 nm, 共 81 个激发波长和 751 个发射波长^[28-29]。测得 1 个样本完整的三维荧光光谱约需要 2.5 min。

1.4 荧光数据预处理

对 300 个留兰香样本采集三维荧光数据,得到 300 个 81 $\times$ 751 的数据矩阵。原始光谱数据会受到外界因素的影响而产生干扰,因此需对原始光谱数据进行预处理。采用三角形内插值法去除散射^[30],用 SG (savitzky-golay) 平滑方法对光谱进行降噪处理^[31]。论文采用的主要方法及流程图如图 1 所示。



注 (Note): SG 为 savitzky-golay, KPCA 为 kernel principal component analysis, KLPP 为 kernel locality preserving projections, FDA 为 fisher discriminant analysis, SVM 为 support vector machine, 下同。

图 1 主要方法及流程图
Fig.1 Main methods and flow chart

1.5 特征提取

1.5.1 数据降维方法

假设存在变量个数为 m , 样本个数为 n 的数据集 $X=[x_1, x_2, \dots, x_n] \in R^{m \times n}$ 。数据集 X 通过非线性映射 $\varphi(x)$, 转换为高维空间数据集 $\Phi(X)=[\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)]$ 。KPCA+KLPP+Wilks Λ 统计量序贯融合算法的目的是通过矩阵 $A \in R^{m \times n}$ 将高维空间数据 $\Phi(X)$ 投影到 $d(d < m)$ 维空间, 将其转换为低维空间数据集 $Y=[y_1, y_2, \dots, y_n]$, 其中

$$Y=A^T\Phi(X)。$$

KPCA 能够将高维数据的全局特征在低维空间保持, 在非线性数据降维任务上广泛应用^[32]。假设存在投影矩阵 A , 则 KPCA 的目标函数 F_{KPCA} 为

$$F_{KPCA}(A) = \max \frac{1}{n} Y Y^T = \max \frac{1}{n} A^T \Phi(X) \Phi^T(X) A \quad (1)$$

式中 n 表示样本个数。

KLPP 算法能够保存数据优化后的邻域结构, 同时

展现数据的流形结构^[33]。KLPP 的目标函数 F_{KLPP} 为

$$F_{\text{KLPP}}(\mathbf{A}) = \min \mathbf{A}^T \Phi(\mathbf{X}) \mathbf{L} \Phi^T(\mathbf{X}) \mathbf{A} \quad (2)$$

式中 $\mathbf{L}$ 为拉普拉斯矩阵。定义核矩阵 $\mathbf{K}_{ij} \in R^{n \times n}$ 。

$$\mathbf{K}_{ij} = \mathbf{K}(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) = \phi^T(x_i) \phi(x_j) \quad (3)$$

选择高斯核函数:

$$\mathbf{K}(x, y) = \exp(-\|x - y\|^2 / \delta) \quad (4)$$

式中 δ 为核窗宽函数, 那么:

$$\begin{aligned} \|\phi(x_i) - \phi(x_j)\|^2 &= (\phi(x_i) - \phi(x_j))^T (\phi(x_i) - \phi(x_j)) \\ &= (\phi^T(x_i) - \phi^T(x_j)) (\phi(x_i) - \phi(x_j)) \\ &= \phi^T(x_i) \phi(x_i) - \phi^T(x_i) \phi(x_j) - \phi^T(x_j) \phi(x_i) \\ &\quad + \phi^T(x_j) \phi(x_j) \\ &= \mathbf{K}_{ii} - \mathbf{K}_{ij} - \mathbf{K}_{ji} + \mathbf{K}_{jj} = 2 - 2\mathbf{K}_{ij} \end{aligned} \quad (5)$$

设中心化的数据为 $\Phi(\bar{\mathbf{X}}) = [\phi(\bar{x}_1), \phi(\bar{x}_2), \dots, \phi(\bar{x}_n)]$, 则 KPCA-KLPP 算法的目标函数为

$$\begin{aligned} F &= F_{\text{KLPP}}(\mathbf{A}) - F_{\text{KPCA}}(\mathbf{A}) = \mathbf{A}^T \Phi(\bar{\mathbf{X}}) \mathbf{L} \Phi^T(\bar{\mathbf{X}}) \mathbf{A} \\ &\quad - \frac{1}{n} \mathbf{A}^T \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}) \mathbf{A} \\ \text{s.t. } \mathbf{A}^T \mathbf{A} &= \mathbf{I} \end{aligned} \quad (6)$$

式中 s. t. 表示约束条件, $\mathbf{I}$ 为单位阵。

将拉格朗日函数引入计算式 (6), 可得到:

$$\begin{aligned} \mathbf{L}(\mathbf{A}, \lambda) &= \text{tr}(\mathbf{A}^T \Phi(\bar{\mathbf{X}}) \mathbf{L} \Phi^T(\bar{\mathbf{X}}) \mathbf{A}) - \frac{1}{n} \mathbf{A}^T \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}) \mathbf{A} \\ &\quad - \lambda (\mathbf{A}^T \mathbf{A} - \mathbf{I}) \end{aligned} \quad (7)$$

式中 $\text{tr}(\cdot)$ 表示矩阵的迹。式 (7) 对 $\mathbf{A}$ 和 λ 分别求导得:

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial \mathbf{A}} &= \Phi(\bar{\mathbf{X}}) \mathbf{L} \Phi^T(\bar{\mathbf{X}}) \mathbf{A} + (\mathbf{A}^T \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}))^T \\ &\quad - \frac{1}{n} \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}) \mathbf{A} - \frac{1}{n} (\mathbf{A}^T \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}))^T - 2\lambda \mathbf{A} \\ &= 2\Phi(\bar{\mathbf{X}}) \mathbf{L} \Phi^T(\bar{\mathbf{X}}) \mathbf{A} - \frac{2}{n} \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}) \mathbf{A} - 2\lambda \mathbf{A} = 0 \end{aligned} \quad (8)$$

$$\frac{\partial \mathbf{L}}{\partial \lambda} = \mathbf{A} \mathbf{A}^T - \mathbf{I} = 0 \quad (9)$$

存在矩阵 $\mathbf{Z}$, 令 $\mathbf{A} = \Phi(\bar{\mathbf{X}}) \mathbf{Z}$ 。则式 (8) 可简化为

$$\Phi(\bar{\mathbf{X}}) \mathbf{L} \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \mathbf{Z} - \frac{1}{n} \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \mathbf{Z} = \lambda \Phi(\bar{\mathbf{X}}) \mathbf{Z} \quad (10)$$

令式 (10) 两边左乘 $\Phi^T(\bar{\mathbf{X}})$ 可得:

$$\begin{aligned} \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \mathbf{L} \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \mathbf{Z} - \frac{1}{n} \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \mathbf{Z} \\ = \lambda \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) \mathbf{Z} \end{aligned} \quad (11)$$

整理可得:

$$\bar{\mathbf{K}} \mathbf{L} \bar{\mathbf{K}} \mathbf{Z} - \frac{1}{n} \bar{\mathbf{K}} \bar{\mathbf{K}} \mathbf{Z} = \lambda \bar{\mathbf{K}} \mathbf{Z} \quad (12)$$

式中 $\bar{\mathbf{K}}$ 是中心化的核矩阵。则 KPCA-KLPP 算法的目标函数式 (6) 最后转化为式 (12)。

Wilks Λ 统计量能够检验不同样本中的各个变量的判别能力^[34], 其计算式为

$$\Lambda = \frac{|D|}{|M|} \quad (13)$$

式中 D 为主成分得分矩阵中每列数据的组内离差, M 为矩阵的总离差。

Λ 值越小表示变量间越有差异。通过求解式 (12) 中矩阵 $\bar{\mathbf{K}} \mathbf{L} \bar{\mathbf{K}} \mathbf{Z} - \frac{1}{n} \bar{\mathbf{K}} \bar{\mathbf{K}} \mathbf{Z}$ 和矩阵 $\bar{\mathbf{K}}$ 的广义特征值得到 $\lambda_1, \lambda_2, \dots, \lambda_n$ 及其对应的特征向量 $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$, 对特征向量组成的矩阵 $\mathbf{T}_1 = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n]$ 构建对应的 Wilks Λ 统计量, 并按照 Wilks Λ 值从小到大进行排序, 从中挑选出前 d 个特征向量组成矩阵 $\mathbf{T}_2 = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_d]$, 那么 $\Phi(\bar{\mathbf{X}})$ 的低维表示为

$$\mathbf{Y} = \mathbf{A}^T \Phi(\bar{\mathbf{X}}) = \mathbf{T}_2 \Phi^T(\bar{\mathbf{X}}) \Phi(\bar{\mathbf{X}}) = \mathbf{T}_2^T \bar{\mathbf{K}} \quad (14)$$

式中 $\mathbf{Y}$ 为低维空间数据集。

KPCA 融入 KLPP 的思想后, 可使投影后的特征空间同时保持原始样本空间的整体结构和局部近邻结构, 再引入 Wilks Λ 统计量进行排序, 能够使后续挑选出的特征波长更具判别能力。

1.5.2 光谱数据特征波长提取及表征

由预处理得到 5 个产地 300 个 81×751 的三维荧光光谱数据矩阵, 先将每个产地的 60 个三维荧光光谱数据矩阵按发射波长合并为 5 个 $4860 (81 \times 60) \times 751$ 的数据矩阵, 再将 5 个数据矩阵按发射波长合并为 1 个 $24300 (81 \times 300) \times 751$ 的数据矩阵^[26]。

分别采用 KPCA、KPCA+KLPP、KPCA+Wilks Λ 统计量、KPCA+KLPP+Wilks Λ 统计量 4 种算法处理合并后的矩阵, 选择累计贡献率达到 99% 的发射波长作为特征发射波长。将选出的特征发射波长按照激发波长合并为 1 个数据矩阵 (相同产地数据放在一起), 再分别用 4 种算法处理该矩阵, 然后选取累计贡献率达到 99% 的激发波长作为特征激发波长^[27]。

将上述 4 种方法得到的特征激发波长从小到大排序, 并将其对应的特征发射波长光谱值首尾相连转换成行向量。最终, 在 300 个样本中, 由 KPCA 与 KPCA+Wilks Λ 统计量方法各得到 1 个 300×320 的特征矩阵作为原始三维荧光数据的表征结果; 由 KPCA+KLPP 与 KPCA+KLPP+Wilks Λ 统计量方法各得到 1 个 300×500 的特征矩阵作为原始三维荧光数据的表征结果。

1.6 鉴别模型输入向量构造

FDA^[35] 作为线性分类方法, 对于多分类问题效果不佳。但可以利用 FDA 进行数据融合, 将不同产地的差异性信息融合生成可分性 FD (fisher discriminant) 变量, 并以其为鉴别模型输入向量, 使鉴别结果更加准确。

将上述三维荧光数据的表征结果用 FDA 算法融合, 可得到融合了特征波长差异性信息的 FD 变量。选择累计判别能力达到 99% 的 FD 变量, 作为鉴别模型的输入向量, 为下一步的鉴别提供更具判别优势的输入向量。

1.7 鉴别模型构建方法

SVM 可以通过映射将原始数据映射到高维空间, 从而解决低维空间中的难题。此外, 在小样本问题的解决上, SVM 具有很好的效果^[36]。

核函数为径向基函数（radial basis function, RBF）的 SVM 无需很大的样本量就可以使 SVM 模型拥有很好的效果^[37]。因此，选择 RBF 作为 SVM 的核函数。为了使 SVM 的鉴别结果更加准确，采用网格搜索法确定最优核函数参数，采用 K 折交叉验证（K-fold cross validation, KCV）法进行验证以获得全局最优解^[38]，减少过拟合现象。

2 结果与分析

2.1 光谱预处理

对采集的 5 个产地留兰香荧光光谱进行预处理，作为示例，图 2 为安徽留兰香的光谱数据预处理图。可以看出，原始荧光数据收到散射的干扰严重，而预处理有效的去除了散射，并且经过 SG 处理后使等高线周围更加平滑。这表示预处理减弱了外部因素的影响，有利于下一步工作的进行。

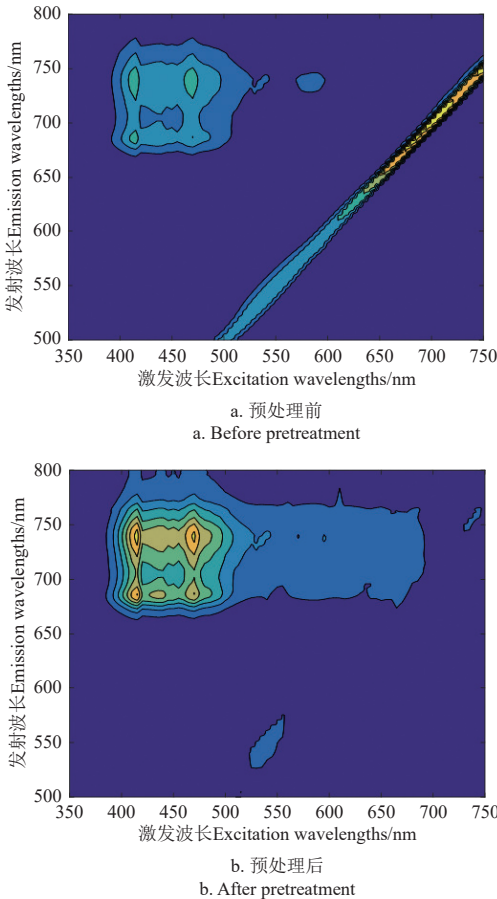


图 2 安徽留兰香预处理前后荧光光谱图
Fig.2 Fluorescence spectra of Anhui spearmint before and after pretreatment

2.2 特征波长提取及表征

利用 5 个产地留兰香的初始三维荧光数据对其进行鉴别是困难的。因此，需要对预处理后的光谱数据进行特征波长的提取，筛选出光谱数据中的有效信息，提高后续产地鉴别的准确率。

对 5 个产地留兰香预处理后的三维荧光数据分别用 KPCA、KPCA+KLPP、KPCA+Wilks Λ 统计量、KPCA+KLPP+Wilks Λ 统计量等 4 种算法按照 1.5.2 中给出的方

法进行特征波长提取，得到的特征波长数见表 2。用 KPCA+KLPP+Wilks Λ 统计量算法得到的最优波长分布图如图 3 所示。

表 2 4 种方法特征波长提取及表征结果
Table 2 Feature wavelength extraction and characterization results of four methods

特征提取方法 Feature extraction methods	特征激发波长数 Feature excitation wavelength number	特征发射波长数 Feature emission wavelength number	特征波长表征结果 Feature wavelength characterization results
KPCA	8	40	300×320
KPCA+KLPP	10	50	300×500
KPCA+Wilks Λ 统计量	8	40	300×320
KPCA+KLPP+Wilks Λ 统计量	10	50	300×500

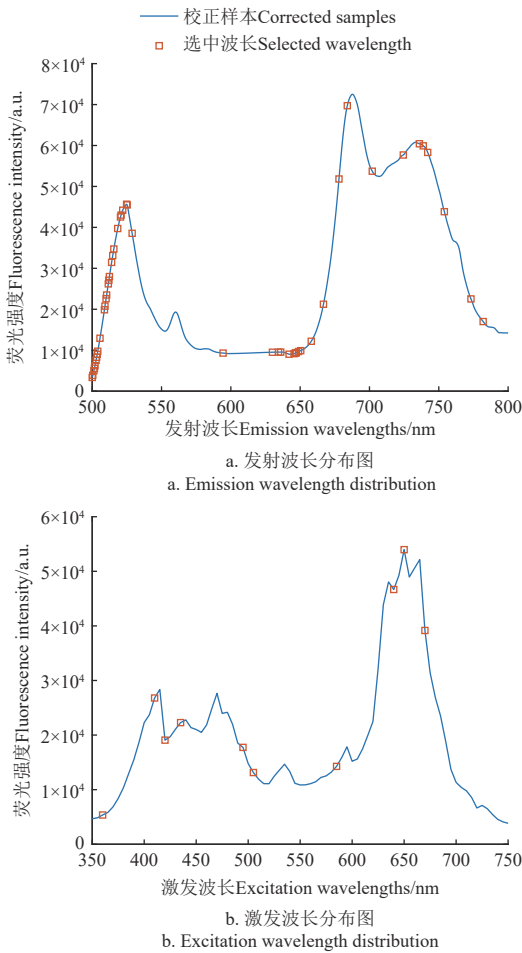


图 3 基于 KPCA+KLPP+Wilks Λ 统计量的特征波长选择分布图
Fig.3 Feature wavelength selection distribution map based on KPCA + KLPP + Wilks Λ statistics

对于每个样本，按特征激发波长从小到大顺序将其对应的特征发射波长的光谱值首尾相连转换成行向量。那么对于 300 个样本，由 KPCA 与 KPCA+Wilks Λ 统计量 2 种方法可各得到 1 个 300×320 的特征矩阵作为所测三维荧光数据的表征结果；由 KPCA+KLPP 与 KPCA+KLPP+Wilks Λ 统计量 2 种方法可各得到 1 个 300×500 的特征矩阵作为所测三维荧光数据的表征结果。

2.3 鉴别模型输入向量构造

采用 FDA 算法融合上述三维荧光数据表征结果中所包含的不同产地的差异性信息, 生成可分性 FD 变量。选择前 4 个累计判别能力达到 99% 的 FD 变量 (每个 FD 变量的判别能力见表 3) 作为 SVM 鉴别模型的输入向量, 进行产地鉴别。

表 3 不同特征波长提取方法对应的前 4 个 FD 变量判别能力
Table 3 Discriminant ability of the first four FD (fisher discriminant) variables corresponding to different feature wavelength methods %

特征提取方法 Feature extraction methods	FD1	FD2	FD3	FD4
KPCA	61.96	19.78	12.31	5.21
KPCA+KLPP	57.60	23.41	11.91	6.18
KPCA+Wilks Λ 统计量	67.07	14.24	10.97	7.02
KPCA+KLPP+Wilks Λ 统计量	82.01	9.54	6.09	2.35

2.4 鉴别模型建立与验证

将挑选出的 4 个 FD 变量作为 SVM 模型的输入向量。4 种不同特征提取方法对应的 4 种模型的产地鉴别结果如图 4~图 7 所示。

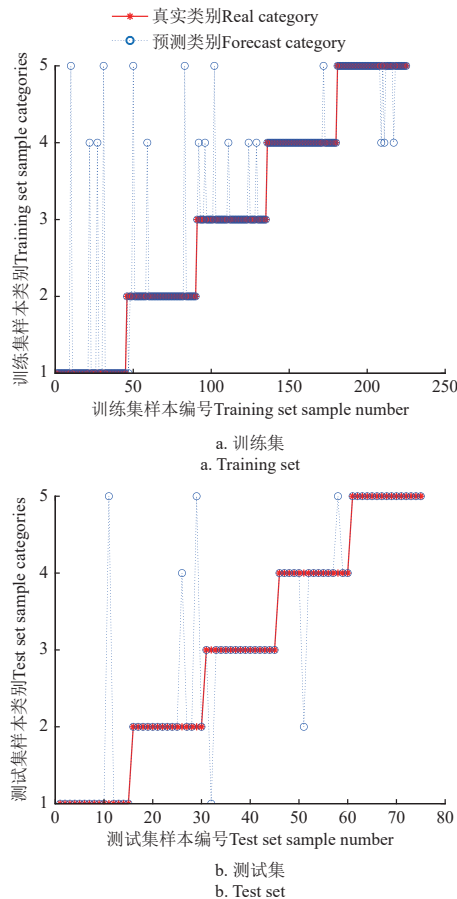


图 4 基于 KPCA 方法的训练集与测试集鉴别结果
Fig.4 Identification results of training set and test set based on KPCA method

由图得, 基于不同特征波长提取方法的 SVM 模型对应的训练集正确率分别为 92.00%、97.78%、95.11%、100%, 测试集正确率分别为 92.00%、96.00%、94.67%、100%。

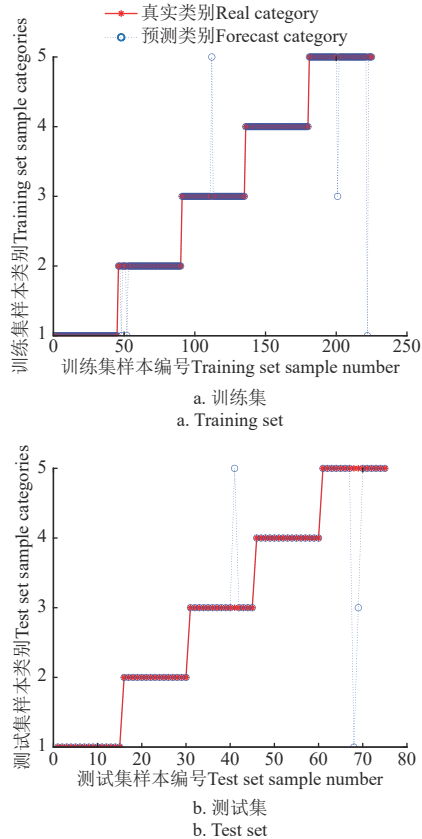


图 5 基于 KPCA+KLPP 方法的训练集与测试集鉴别结果
Fig.5 Identification results of training set and test set based on KPCA+KLPP method

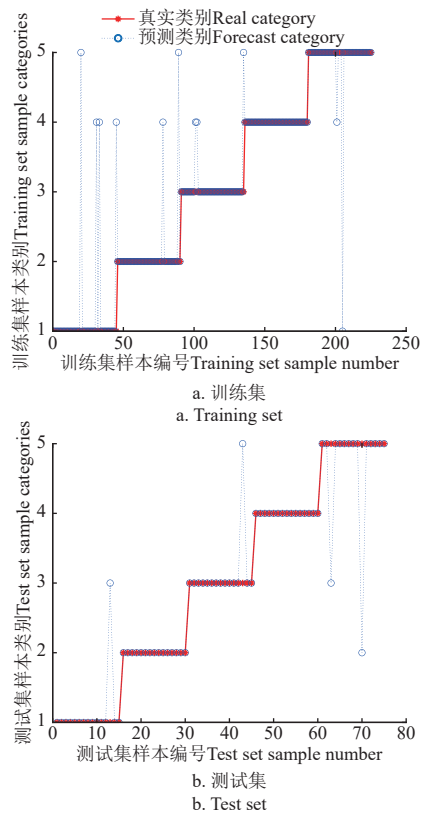


图 6 基于 KPCA+Wilks Λ 统计量方法的训练集与测试集鉴别结果
Fig.6 Identification results of training set and test set based on KPCA+Wilks Λ statistic method

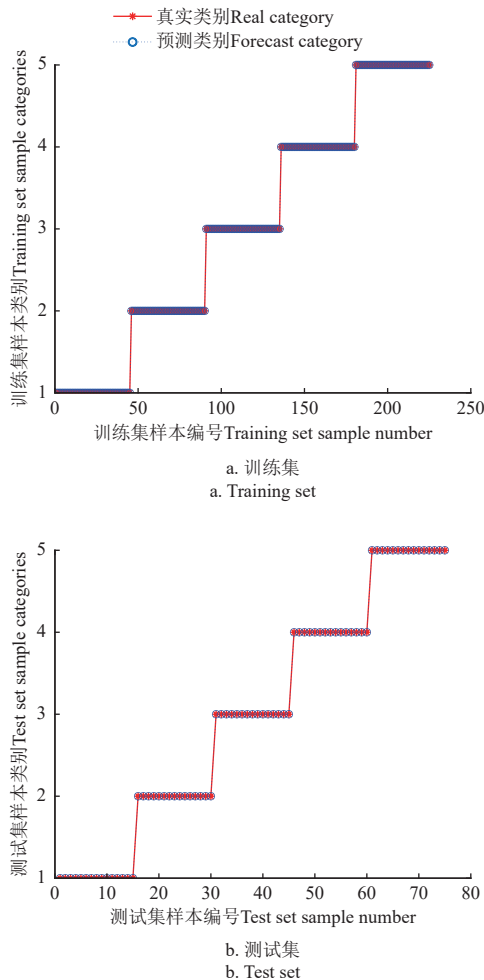


图 7 基于 KPCA+KLPP+Wilks Λ 统计量方法的训练集与测试集鉴别结果

Fig.7 Identification results of training set and test set based on KPCA+KLPP+Wilks Λ statistic method

对比分析可知,使用 KPCA+KLPP 作为特征提取方法比仅使用 KPCA 方法测试集正确率从 92.00% 提高至 96.00%。这是因为 KPCA 方法仅保留了数据的全局方差信息,忽略了局部信息;而 KLPP 可以最大程度地使原始数据中的局部信息得以保留。因此,KPCA+KLPP 方法在保证全局方差信息不变的同时保持了样本数据的局部信息。使用 KPCA+Wilks Λ 统计量作为特征提取方法比仅使用 KPCA 方法测试集正确率从 92.00% 提高至 94.67%,这说明使用 KPCA+Wilks Λ 统计量方法比仅使用 KPCA 方法筛选出的变量判别能力更强。使用 KPCA+KLPP+Wilks Λ 统计量作为特征提取方法比仅使用 KPCA 方法测试集正确率从 92.00% 提高至 100%。由此可见,KLPP 和 Wilks Λ 统计量均可在一定程度上对 KPCA 进行改进完善,3 者序贯融合可以使整体特征提取能力明显提升,能更好的去除光谱数据的冗余信息,提高鉴别的正确率。因此,KPCA+KLPP+Wilks+FDA+SVM 序贯融合的方法可以作为留兰香产地鉴别实施的有效策略。

3 结 论

研究选取河南、安徽、北京、上海和广东 5 个产地

的留兰香进行产地鉴别。首先,在采集留兰香三维荧光光谱数据后对原始光谱数据进行预处理。然后,采用 KPCA (kernel principal component analysis), KPCA+KLPP (kernel locality preserving projections), KPCA+Wilks Λ 统计量, KPCA+KLPP+Wilks Λ 统计量 4 种方法进行特征波长提取。接着,用 FDA (fisher discriminant analysis) 构建鉴别模型输入向量。最后,通过构建 SVM (support vector machine) 模型实现了 5 个产地留兰香的鉴别。这为后续利用三维荧光光谱开展留兰香重要组分量化分析奠定了基础。研究主要结论如下:

1) 提出了一种 KPCA+KLPP+Wilks Λ 统计量序贯融合的三维荧光数据特征波长提取策略。通过对 4 种不同特征波长提取策略进行对比后发现,该特征波长提取策略能够更有效地减少光谱数据的冗余信息,表征原始三维荧光数据的可分性信息,提高产地鉴别的正确率。

2) 用 FDA 算法可融合不同产地的差异性信息,得到可分性 FD 变量。选择前 4 个累计判别能力达到 99% 的 FD (fisher discriminant) 变量,以其作为鉴别模型输入向量,降低了鉴别模型构建的复杂度。

3) 借助于 KPCA+KLPP+Wilks Λ 统计量序贯融合的三维荧光数据特征波长提取策略和 FDA 生成的可分性 FD 变量,成功构建了 SVM 鉴别模型,鉴别正确率可达 100%,实现了 5 个产地留兰香正确鉴别。

[参 考 文 献]

- [1] 李民生,崔宇宏,王爱娜,等.种植、野生薄荷原药材调研考察及分析[J].中国药事,2019,33(4): 435-445.
LI Minsheng, CUI Yuhong, WANG Aina, et al. Investigation and analysis into cultivated and wild mentha haplocalyx briaq[J]. Chinese Pharmaceutical Affairs, 2019, 33(4): 435-445. (in Chinese with English abstract)
- [2] 韩战强,霍磊,李鹏伟,等.留兰香的功能作用及在动物生产中的应用研究进展[J].中国饲料,2021(23): 20-23.
HAN Zhanqiang, HUO Lei, LI Pengwei, et al. The function of mentha spicata and its application in animal production[J]. China Feed, 2021(23): 20-23. (in Chinese with English abstract)
- [3] 王莹莹,陈鑫沛,董颖,等.留兰香调味油的研制[J].中国调味品,2022,47(4): 117-121.
WANG Yingying, CHEN Xinpei, DONG Ying, et al. The preparation of Mentha spicata L. seasoning oil[J]. China Condiment, 2022, 47(4): 117-121. (in Chinese with English abstract)
- [4] SAM H V, AMIN N, DERICK M, et al. Prediction of essential oil content in spearmint (Mentha spicata) via near-infrared hyperspectral imaging and chemometrics[J]. Scientific reports, 2023, 13(1): 4261.
- [5] 卢先勇,何文绚,陈昊聪,等.基于植物油三维荧光光谱的茶油鉴定模型[J].光谱学与光谱分析,2020,40(12): 3854-3859.
LU Xianyong, HE Wenxuan, CHEN Haocong, et al. Identification model of camellia oil based on 3D fluorescence spectra of vegetable oils[J]. Spectroscopy and Spectral Analysis, 2020, 40(12): 3854-3859. (in Chinese with English abstract)

- [6] XUE J T, SHI Y L, YE L M, et al. Near-infrared spectroscopy for rapid and simultaneous determination of five main active components in rhubarb of different geographical origins and processing[J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2018, 205: 419-427.
- [7] 樊凤杰, 轩凤来, 白洋, 等. 基于三维荧光光谱特征的中草药性模式识别研究[J]. *光谱学与光谱分析*, 2020, 40(6): 1763-1768.
FAN Fengjie, XUAN Fenglai, BAI Yang, et al. Pattern recognition of traditional Chinese medicine property based on three-dimensional fluorescence spectrum characteristics[J]. *Spectroscopy and Spectral Analysis*, 2020, 40(6): 1763-1768. (in Chinese with English abstract)
- [8] 康超娣, 王守伟, 张颖颖, 等. 液相色谱-串联质谱法对牛肉中掺假成分的相对定量分析[J]. *食品科学*, 2022, 43(4): 270-276.
KANG Chaodi, WANG Shouwei, ZHANG Yingying, et al. Quantitative analysis of adulterants in beef by liquid chromatography-tandem mass spectrometry[J]. *Food Science*, 2022, 43(4): 270-276. (in Chinese with English abstract)
- [9] DENG C H, YAO N, WANG B, et al. Development of microwave-assisted extraction followed by headspace single-drop microextraction for fast determination of paeonol in traditional Chinese medicines[J]. *Journal of chromatography, A: Including electrophoresis and other separation methods*, 2006, 1103(1): 15-21.
- [10] REGUEIRO J, SÁNCHEZ-GONZÁLEZ C, VALLVERDÚ-QUERALT A, et al. Comprehensive identification of walnut polyphenols by liquid chromatography coupled to linear ion trap-Orbitrap mass spectrometry[J]. *Food Chemistry*, 2014, 152: 340-348.
- [11] 张伟, 张勇, 尹震花, 等. 超声辅助离子液体微萃取-反相液相色谱法测定人工蛹虫草活性成分[J]. *食品科学*, 2017, 38(8): 179-183.
ZHANG Wei, ZHANG Yong, YIN Zhenhua, et al. Determination of the active ingredients in cultured cordyceps militaris by ultrasound-assisted ionic liquid microextraction coupled with reversed phase liquid chromatography[J]. *Food Science*, 2017, 38(8): 179-183. (in Chinese with English abstract)
- [12] 吴子阳, 何菲, 李贺贺, 等. 采用固相萃取结合高效液相色谱测定清香型白酒中酚类活性化合物[J]. *分析化学*, 2020, 48(10): 1400-1408.
WU Ziyang, HE Fei, LI Hehe, et al. Determination of phenolic active compounds in light-aroma Baijiu by solid phase extraction-high performance liquid chromatography[J]. *Chinese Journal of Analytical Chemistry*, 2020, 48(10): 1400-1408. (in Chinese with English abstract)
- [13] MA J M, FAN S F, YANG L Q, et al. Rapid screening of 420 pesticide residues in fruits and vegetables using ultra high performance liquid chromatography combined with quadrupole-time of flight mass spectrometry[J]. *Food Science and Human Wellness*, 2023, 12(4): 1064-1070.
- [14] CHEN J Y, CHEN X W, LIN Y Y, et al. Authentication of dark brown sugars from different processing using three-dimensional fluorescence spectroscopy[J]. *LWT-Food Science and Technology*, 2021, 150: 111959.
- [15] 郭怡宁, 杨振清, 朱炳祺. 三维荧光技术在中药质量评价中应用的研究进展[J]. *中国现代中药*, 2021, 23(11): 2025-2030.
GUO Yining, YANG Zhenqing, ZHU Bingqi. Application of three-dimensional fluorescence technique for quality evaluation of traditional Chinese medicine: A review[J]. *Modern Chinese Medicine*, 2021, 23(11): 2025-2030. (in Chinese with English abstract)
- [16] 张轩, 高磊, 刘安琴, 等. 宁夏枸杞果实的三维荧光光谱特性研究[J]. *中药材*, 2017, 40(5): 1146-1149.
- [17] 鹿鸣. 基于光谱技术的粳稻稻种鉴别研究[D]. 长春: 长春理工大学, 2023.
LU Ming. Identification of Japonica Rice Seeds Based on Spectral Technology[D]. Changchun: Changchun University of Science and Technology, 2023. (in Chinese with English abstract)
- [18] CHENG P F, WANG S C, ZHU Y P, et al. Application of three-dimensional fluorescence spectroscopy in smart agriculture — detection of oil pollutants in water[J]. *International Journal of Pattern Recognition and Artificial Intelligence*, 2023, 37(3): 2355004.
- [19] 孔德明, 陈红杰, 陈晓玉, 等. 三维荧光光谱结合稀疏主成分分析和支持向量机的油类识别方法研究[J]. *光谱学与光谱分析*, 2021, 41(11): 3474-3479.
KONG Deming, CHEN Hongjie, CHEN Xiaoyu, et al. Research on oil identification method based on three-dimensional fluorescence spectroscopy combined with sparse principal component analysis and support vector machine[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(11): 3474-3479. (in Chinese with English abstract)
- [20] 吴晓婷, 闫德勤. 数据降维方法分析与研究[J]. *计算机应用研究*, 2009, 26(8): 2832-2835.
WU Xiaoting, YAN Deqin. Analysis and research on method of data dimensionality reduction[J]. *Application Research of Computers*, 2009, 26(8): 2832-2835. (in Chinese with English abstract)
- [21] MULLER M, JIMENEZ J, ANTONINI M, et al. Combining chemical sequential extractions with 3D fluorescence spectroscopy to characterize sludge organic matter[J]. *Waste Management*, 2014, 34(12): 2572-2580.
- [22] 李小占, 马本学, 喻国威, 等. 基于深度学习与图像处理的哈密瓜表面缺陷检测[J]. *农业工程学报*, 2021, 37(1): 223-232.
LI Xiaozhan, MA Benxue, YU Guowei, et al. Surface defect detection of Hami melon using deep learning and image processing[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2021, 37(1): 223-232. (in Chinese with English abstract)
- [23] 周博, 朱文魁, 王赵改, 等. 基于太赫兹时域光谱技术的烟草组分识别[J]. *农业工程学报*, 2022, 38(10): 310-316.
ZHOU Bo, ZHU Wenkui, WANG Zhaogai, et al. Identification of tobacco materials based on terahertz time-domain spectroscopy[J]. *Transactions of the Chinese Society of*

- Agricultural Engineering (Transactions of the CSAE)*, 2022, 38(10): 310-316. (in Chinese with English abstract)
- [24] 殷勇, 田先亮. 基于 PCA 与 Wilks 准则的电子鼻酒类鉴别方法研究[J]. *仪器仪表学报*, 2007, 28(5): 849-852.
YIN Yong, TIAN Xianliang. Study of drink identification method using electronic nose based on PCA and Wilks rule[J]. *Chinese Journal of Scientific Instrument*, 2007, 28(5): 849-852. (in Chinese with English abstract)
- [25] 张学磊, 徐进学, 祝朋艳. 基于改进核局部保持投影的故障检测方法研究[J]. *计算机应用与软件*, 2022, 39(5): 115-124, 131.
ZHANG Xuelei, XU Jinxue, ZHU Pengyan. Research on fault detection method based on modified kernel locality preserving projections[J]. *Computer Applications and Software*, 2022, 39(5): 115-124, 131. (in Chinese with English abstract)
- [26] 于慧春, 李迎, 殷勇, 等. 苹果贮藏室气体 3D 荧光特征信息小波包表征与腐败预警[J]. *农业机械学报*, 2022, 53(5): 392-399.
YU Huichun, LI Ying, YIN Yong, et al. Wavelet packet characterization of 3D fluorescence characteristic Information of gas in storage room and early warning method of apple spoilage[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2022, 53(5): 392-399. (in Chinese with English abstract)
- [27] 任永杰, 殷勇, 于慧春, 等. 三维荧光光谱融合小波包分解融合 Fisher 判别分析及支持向量机识别紫苏[J]. *食品科学*, 2024, 45(1): 198-203.
REN Yongjie, YIN Yong, YU Huichun, et al. Identification of perilla based on three-dimensional fluorescence spectra using wavelet packet decomposition, Fisher discriminant analysis and support vector machine[J]. *Food Science*, 2024, 45(1): 198-203. (in Chinese with English abstract)
- [28] 刘瑞敏, 殷勇, 于慧春, 等. 多变量统计分析融合小波包能量提取 3D 荧光特征信息监测黄瓜贮藏品质[J]. *光谱学与光谱分析*, 2023, 43(9): 2967-2973.
LIU Ruimin, YIN Yong, YU Huichun, et al. Extraction of 3D fluorescence feature information based on multivariate statistical analysis coupled with wavelet packet energy for monitoring quality change of cucumber during storage[J]. *Spectroscopy and Spectral Analysis*, 2023, 43(9): 2967-2973. (in Chinese with English abstract)
- [29] 李建盟, 殷勇, 于慧春, 等. 番茄贮藏气氛 3D 荧光特征波长小波包选择及腐败预警方法[J]. *食品科学*, 2022, 43(21): 63-69.
LI Jianmeng, YIN Yong, YU Huichun, et al. Feature wavelength selection of three-dimensional fluorescence data of tomato storage room gas based on wavelet packet decomposition for early warning of its spoilage[J]. *Food Science*, 2022, 43(21): 63-69. (in Chinese with English abstract)
- [30] 孔德明, 李雨蒙, 崔耀耀, 等. 海面溢油三维荧光光谱消除瑞利散射方法的研究[J]. *光谱学与光谱分析*, 2020, 40(9): 2791-2797.
KONG Deming, LI Yumeng, CUI Yaoyao, et al. Correction methods of rayleigh scattering of three-dimensional fluorescence spectra of spilled oil on sea[J]. *Spectroscopy and Spectral Analysis*, 2020, 40(9): 2791-2797. (in Chinese with English abstract)
- [31] 张晓东, 段朝晖, 毛罕平, 等. 利用太赫兹光谱技术构建番茄水分胁迫状态检测模型[J]. *农业工程学报*, 2021, 37(15): 121-128.
ZHANG Xiaodong, DUAN Chaohui, MAO Hanping, et al. Tomato water stress state detection model by using terahertz spectroscopy technology[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2021, 37(15): 121-128. (in Chinese with English abstract)
- [32] 乡立, 段炼, 陈菁, 等. 基于 WEKPCA 与深度学习的短期电力负荷预测[J]. *电力信息与通信技术*, 2020, 18(10): 34-41.
XIANG Li, DUAN Lian, CHEN Jing, et al. Short-term power load forecasting based on WEKPCA and deep learning[J]. *Electric Power Information and Communication Technology*, 2020, 18(10): 34-41. (in Chinese with English abstract)
- [33] 郭金玉, 王霞, 李元. 基于 MEWMA 的自适应 KLPP 的非线性过程故障检测[J/OL]. *控制理论与应用*, [2024-03-01]. <http://kns.cnki.net/kcms/detail/44.1240.TP.20230511.1653.030.html>.
GUO Jinyu, WANG Xia, LI Yuan. Fault detection of nonlinear process based on adaptive KLPP of MEWMA[J/OL]. *Control Theory & Applications*, [2024-03-01]. <http://kns.cnki.net/kcms/detail/44.1240.TP.20230511.1653.030.html>. (in Chinese with English abstract)
- [34] 李孟丽, 殷勇, 袁云霞, 等. 香蕉贮藏气体 3D 荧光表征特征选择及早期腐败预警初探[J]. *光谱学与光谱分析*, 2021, 41(2): 558-564.
LI Mengli, YIN Yong, YUAN Yunxia, et al. Feature selection of 3D fluorescence data based on storage room gas and preliminary early warning of banana spoilage[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(2): 558-564. (in Chinese with English abstract)
- [35] 殷勇, 王燕芳, 葛飞, 等. 基于空载数据的鉴别食醋电子鼻信号漂移校正方法[J]. *农业工程学报*, 2019, 35(17): 293-300.
YIN Yong, WANG Yanfang, GE Fei, et al. E-nose information drift correction method for identifying vinegar based on no-load data[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2019, 35(17): 293-300. (in Chinese with English abstract)
- [36] 潘春华, 肖德琴, 林探宇, 等. 基于 SVM 和区域生长结合算法的南方主要蔬菜害虫分类识别 (英文) [J]. *农业工程学报*, 2018, 34(8): 192-199.
PAN Chunhua, XIAO Deqin, LIN Tanyu, et al. Classification and recognition for major vegetable pests in Southern China using SVM and region growing algorithm[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2018, 34(8): 192-199. (in Chinese with English abstract)
- [37] HUANG W C, LIU H Y, ZHANG Y, et al. Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM[J]. *Applied Soft Computing*, 2021, 109: 107541.
- [38] 刘津彤, 张岚泽, 姜红, 等. 差分拉曼光谱结合支持向量

机模型对便签纸的鉴别分析[J]. 化学通报, 2022, 85(2): 259-263, 246.
LIU Jintong, ZHANG Lanze, JIANG Hong, et al.

Identification of note paper by differential raman spectroscopy based on support vector machine[J]. Chemistry, 2022, 85(2): 259-263, 246. (in Chinese with English abstract)

Spearmint three-dimensional fluorescence data feature extraction and identification based on KPCA and KLPP and Wilks statistics

YIN Yong , XU Feifan , YU Huichun , YUAN Yunxia

(College of Food and Bioengineering, Henan University of Science and Technology, Luoyang 471000, China)

Abstract: This study aims to rapidly and accurately identify the origin of spearmint. A feature wavelength extraction was proposed to combine the sequential fusion of Kernel Principal Component Analysis (KPCA) with Kernel Locality Preserving Projections (KLPP) and Wilks' Λ statistics. Effective discrimination was realized among the five origins of spearmint. Firstly, the three-dimensional fluorescence data was collected from the 300 spearmint samples in five locations. The interferences of Rayleigh and Raman scattering were removed from the original spectra by triangular interpolation. Savitzky-Golay (SG) was used to smooth and then eliminate the noise. Secondly, the preprocessed three-dimensional fluorescence spectrum matrix was merged into the data matrix, according to the emission wavelength. KPCA, KPCA + KLPP, KPCA + Wilks Λ statistics, and KPCA + KLPP + Wilks Λ statistics were used to extract the feature wavelengths of the three-dimensional fluorescence spectrum matrix after preprocessing. Among them, KPCA and KPCA + Wilks Λ statistics were used to extract 40 wavelengths of feature emission and eight wavelengths of feature excitation, respectively. KPCA+KLPP and KPCA + KLPP + Wilks Λ statistics were to extract 50 wavelengths of feature emission and 10 wavelengths of feature excitation, respectively. Thirdly, the spectral values of feature emission wavelength were converted into row vectors, according to the wavelength order of feature excitation. Four 300-row spectral matrices of feature wavelength were then obtained from the 300 samples. Furthermore, Fisher discriminant analysis (FDA) was applied to the spectral matrix of feature wavelengths for data separable fusion for the separable fisher discriminant (FD) variables. The first four FD variables were selected with a cumulative discriminant ability of 99% as the input vectors of the discriminant model for the spearmint origins. Finally, the support vector machine (SVM) was used to analyze the four FD variables. The FDA + SVM identification model showed that there was a four-feature wavelength extraction. The correct rates for the training set were 92.00%, 97.78%, 95.11%, and 100%, respectively, and the correct rates for the test set were 92.00%, 96.00%, 94.67%, and 100%, respectively. The results show that both KLPP and Wilks Λ statistics improved the performance of KPCA. The sequential fusion significantly improved the overall feature extraction, in order to better remove the redundant information of spectral data for the high correct of identification. Therefore, the sequential fusion of KPCA + KLPP + Wilks Λ statistics + FDA + SVM can be expected to effectively identify the origin of spearmint. The extraction of feature wavelength with sequential fusion of KPCA + KLPP + Wilks Λ statistics can effectively reduce the redundancy of three-dimensional fluorescence spectral data for the informative features of the original fluorescence data. Better identification was achieved in the five kinds of spearmint origins. In addition, these research findings can lay a foundation for the subsequent quantitative analysis of the important components in spearmint by three-dimensional fluorescence spectroscopy.

Keywords: fluorescence spectrum; discriminant analysis; model; spearmint; kernel principal component analysis