基于高效液相色谱指纹图谱结合化学计量学及机器学习的黑茶产地识别

王贞红[1#], 韩沅汐[2#], 张立友[1], 叶永祥[2], 魏丽萍[1], 李 梁[2*]

(1. 西藏农牧学院资源与环境学院, 西藏农牧学院茶产业工程中心, 林芝 860000; 2. 西藏农牧学院食品科学学院, 高原特色农产品研发中心, 西藏特色农牧资源研发协同创新中心, 林芝 860000)

摘 要: 目的 建立广西、湖南、四川、陕西和西藏产地黑茶高效液相色谱(high performance liquid chromatography, HPLC)指纹图谱, 并结合化学计量学与机器学习对黑茶进行产地识别研究。**方法** 采用高效液相色谱-二极管阵列检测器(high performance liquid chromatography-diode array detector, HPLC-DAD)检测 48 份不同产地黑茶的化学成分, 并建立指纹图谱; 利用没食子酸、表没食子儿茶素、表儿茶素没食子酸酯、表儿茶素、儿茶素、咖啡碱和表没食子儿茶素没食子酸酯 7 种对照品对图谱共有峰进行指认; 结合化学计量学和不同机器学习算法建立黑茶产地识别模型, 并使用准确率、精确率、召回率及 F_1 分数作为机器学习产地识别模型的评价指标。**结果** 黑茶指纹图谱共识别出 8 个共有峰, 指认其中 7 个成分; 基于指纹图谱 8 个共有峰峰面积建立的化学计量学和机器学习的产地识别模型中显示, 偏最小二乘法-判别分析模型能识别部分产地黑茶, 并筛选出 4 个差异标志物, 其预测准确率为 54.2%, 逻辑回归(logistic regression, LR)、反向传播神经网络(back propagation neural network, BPNN)、支持向量机(support vector machine, SVM)、随机森林(random forest, RF)和决策树(decision tree, DT)算法模型预测准确率分别为 66.7%、90.0%、90.0%、80.0%和 80.0%, 其中, SVM 模型的产地识别效果最好。**结论** 我国不同产地黑茶化学成分含量存在一定差异, HPLC 指纹图谱结合 SVM 能够较好对黑茶产地进行溯源研究。

关键词: 黑茶; 高效液相色谱; 化学指纹; 产地溯源

Identification of dark tea origin based on high performance liquid chromatography fingerprint combined with chemometrics and machine learning

WANG Zhen-Hong[1#], HAN Yuan-Xi[2#], ZHANG Li-You[1], YE Yong-Xiang[2], WEI Li-Ping[1], LI Liang[2*]

(1. *Resources & Environment College*, *Tibet Agriculture & Animal Husbandry University*; *Tea Industry Engineering Center of Tibet Agriculture and Animal Husbandry University*, *Nyingchi* 860000, *China*; 2. *Food Science College*, *Tibet Agriculture & Animal Husbandry University*; *Research & Development Center of Agricultural Products with Tibetan Plateau Characteristics*; *Provincial and Ministerial Co-founded Collaborative Innovation Center for Research & Development in Tibet Characteristic Agricultural and Animal Husbandry Resources*, *Nyingchi* 860000, *China*)

基金项目: 西藏自治区中央引导地方项目(XZ202201YD0038C)、国家自然科学基金项目(U21A20232)、西藏自治区重点研发专项(XZ202001ZY0035N)

Fund: Supported by the Central Guidance on Local Science and Technology Development Fund of Tibet Autonomous Region (XZ202201YD0038C), the National Natural Science Foundation of China (U21A20232), and the Tibet Key Point Research and Invention Program (XZ202001ZY0035N)

#王贞红、韩沅汐为共同第一作者

#WANG Zhen-Hong and HAN Yuan-Xi are Co-first Authors

***通信作者:** 李梁, 硕士, 副教授, 主要研究方向为高原特色农产品功能成分挖掘与产地溯源。E-mail: jwllok@sina.com

***Corresponding author:** LI Liang, Master, Associate Professor, Food Science College, Tibet Agriculture & Animal Husbandry University, Nyingchi 860000, China. E-mail: jwllok@sina.com

ABSTRACT: Objective　To construct the high performance liquid chromatography (HPLC) fingerprints of dark tea in the five major producing areas of Guangxi, Hunan, Sichuan, Shaanxi, and Tibet, and identify the producing areas based on chemometrics and machine learning. **Method**　A high performance liquid chromatography-diode array detector (HPLC-DAD) method was utilized to analyze the chemical components of 48 different origins of dark tea and establish a fingerprint profile. Seven kinds of reference standards including gallic acid, epicatechin, epicatechin gallate, catechin, theaflavin, caffeine, and epigallocatechin gallate were used to identified the peak. A combination of chemometrics and various machine learning algorithms were employed to establish models for the identification of dark tea origins. Accuracy, precision, recall, and F_1 score were used as evaluation metrics for the machine learning models. **Results**　The fingerprint profile of dark tea identified a total of 8 common peaks, with 7 components being identified. Based on the peak areas of the 8 common peaks in the fingerprint profile, the chemometrics and machine learning models for the origin identification were established. The partial least squares-discriminant analysis model was able to identify some origins of dark tea and identified 4 differential markers, with a prediction accuracy of 54.2%. The logistic regression (LR), back propagation neural network (BPNN), support vector machine (SVM), random forest (RF), and decision tree (DT) algorithms achieved prediction accuracy of 66.7%, 90.0%, 90.0%, 80.0%, and 80.0%, respectively. The evaluation indicated that the SVM model had the best performance for the origin identification. **Conclusion**　The chemical components of dark tea from different producing areas in China are slightly different, and HPLC fingerprints combined with SVM can better trace the producing area of dark tea.

KEY WORDS: dark tea; high performance liquid chromatography; chemical fingerprint; geographical origin traceability

0　引　言

黑茶是我国六大茶类之一[1], 其以成熟度较高的夏秋茶叶为原料, 经杀青、揉捻、渥堆和干燥 4 道工序制成[2], 因其独特的风味和品质, 以及具有抗氧化[3−4]、降血脂[5−6]、抑制肥胖[7−8]和调节肠胃[9−10]等功能, 深受国内外消费者喜爱。我国黑茶主要包括安徽六安黑茶、湖南安化黑茶、四川边茶、云南普洱熟茶、广西六堡茶、陕西茯茶和西藏黑茶等, 部分产地黑茶已成为国内外著名的地理标志产品。因黑茶价格由于原料和加工工艺的差异而大相径庭, 一些商贩为追求经济利益选择伪造黑茶产地标签[11], 导致市面黑茶质量安全难以保证, 不仅严重侵害消费者权益和健康, 还严重损害优质黑茶品牌的商誉, 影响我国黑茶产业健康发展, 故开发一种稳定可靠的黑茶产地识别方法, 对维护其贸易公平具有积极的推动作用。

传统的黑茶产地溯源方法依赖于感官评定, 这要求审评者具备扎实的评茶知识技能和规范的茶叶感官评审操作[12], 但不具备专业技能的普通消费者较难识别伪劣产品。近年来, 已有研究者通过现代化学分析技术对黑茶产地溯源进行研究, 方法主要包括成分分析技术(气相色谱-质谱法[13]、液相色谱-质谱法[14]、液相色谱法[15])、元素分析技术(稳定同位素质谱法[16]、电感耦合等离子体质谱法[17])和光谱技术(红外光谱法[18−19]、核磁共振法[20])等, 其中高效液相色谱法(high performance liquid chromatography, HPLC)具有重复性好、精确度高及试剂用量少等特点[21], 已用于各类特色农产品的产地溯源研究中[22−24]。胡燕[25]利用 HPLC 结合聚类分析对来自我国 6 个不同产地的黑茶样品的共有峰进行了初步区分, 常睿等[26]基于生化成分并使用主成分分析(principal component analysis, PCA)和系统聚类分析探讨来自湖南、湖北、四川和云南的黑茶样本发现, 仅云南熟普样品聚为一类。白秀芝等[27]基于 HPLC 指纹图谱分别采用 PCA、正交偏最小二乘法-判别分析(orthogonal partial least squares-discriminant analysis, OPLS-DA)和随机森林(random forest, RF)识别湖南和非湖南黑茶, 结果显示 PCA 不能有效区分, 而 OPLS-DA 和 RF 均获得了较好的预测结果。综上可知, HPLC 可有效鉴识各产地黑茶间的品质特征差异, 但多采用聚类分析和 PCA 等无监督方法进行产地识别[28], 其识别准确率有待提升, 且对于多产地多变量的黑茶识别效果需进一步验证。

机器学习是一种在经验学习中改善具体算法性能的方法, 具有分类准确率高、预测能力强等特点, 近年来已逐步应用于产地溯源研究中, 并取得较好的效果。CUI 等[29]将核磁共振波谱技术结合多种机器学习算法应用于 7 个产区红茶的产地识别, 采用 RF、支持向量机(support vector machine, SVM)、线性判别分析和 K-最近邻 4 种模型的整体准确率分别为 92.7%、91.8%、86.3%、86.3%。YUN 等[30]探究了顶空-气相色谱法结合 K-近邻和 RF 两种机器学习算法对来自 10 个地区的 306 份红茶进行产地识别的可行性, 结果显示 K-近邻和 RF 模型的预测正确率高达 100.0%和 97.0%。JIN 等[31]利用了近红外光谱技术和极限学习机算法识别对太平猴魁绿茶的产地来源, 获得了 95.3%的预测准确率。以上研究表明机器学习在产地溯源领域中具有巨大潜力, 然

而经文献调研发现，目前采用机器学习的茶叶产地溯源研究主要集中于利用光谱技术对红茶、绿茶及乌龙茶品种和产地的识别[32–33]，尚无关于高效液相色谱法结合机器学习针对我国黑茶产地溯源的研究报道。

鉴于此，本研究采用高效液相色谱-二极管阵列检测器(high performance liquid chromatography-diode array detector, HPLC-DAD)构建我国 5 个地区的 48 份黑茶样品的指纹图谱，对其中的共有峰成分进行指认并分析，利用化学计量学 PLS-DA 及机器学习逻辑回归(logistic regression, LR)、反向传播神经网络(back propagation neural network, BPNN)、SVM、RF 和决策树(decision tree, DT)算法建立黑茶产地识别模型，为我国黑茶产地溯源研究提供参考依据。

1 材料与方法

1.1 材料与试剂

样品取自我国 5 个黑茶主产地，其中 S1~S5 为广西黑茶，S6~S18 为湖南黑茶，S19~S25 为四川黑茶，S26~S30 为陕西黑茶，S31~S48 为西藏黑茶，如表 1 所示为 48 份黑茶样品来源信息。

表 1 48 份黑茶样品来源信息

Table 1 Source information of 48 black tea samples

编号	产地	生产年份	编号	产地	生产年份
S1	广西	2018	S25	四川	2017
S2	广西	2019	S26	陕西	2018
S3	广西	2020	S27	陕西	2019
S4	广西	2021	S28	陕西	2020
S5	广西	2022	S29	陕西	2021
S6	湖南	2012	S30	陕西	2019
S7	湖南	2013	S31	西藏	2022
S8	湖南	2014	S32	西藏	2022
S9	湖南	2017	S33	西藏	2022
S10	湖南	2013	S34	西藏	2022
S11	湖南	2018	S35	西藏	2022
S12	湖南	2018	S36	西藏	2022
S13	湖南	2019	S37	西藏	2022
S14	湖南	2019	S38	西藏	2022
S15	湖南	2021	S39	西藏	2022
S16	湖南	2021	S40	西藏	2022
S17	湖南	2021	S41	西藏	2022
S18	湖南	2022	S42	西藏	2022
S19	四川	2016	S43	西藏	2022
S20	四川	2017	S44	西藏	2022
S21	四川	2017	S45	西藏	2019
S22	四川	2018	S46	西藏	2020
S23	四川	2019	S47	西藏	2021
S24	四川	2021	S48	西藏	2017

没食子酸(纯度≥97.7%)、表没食子儿茶素(纯度≥98.0%)、表儿茶素没食子酸酯(纯度≥98.0%)、表儿茶素(纯度≥99.4%)、儿茶素(纯度≥98.0%)、咖啡碱(纯度≥95.0%)、表没食子儿茶素没食子酸酯(纯度≥98.0%)对照品(上海安谱璀世标准技术服务有限公司)；甲醇(色谱纯，德国默克公司)；乙腈、乙酸(色谱纯，天津科密欧化学试剂有限公司)。

1.2 仪器与设备

安捷伦 1260 Infinity 高效液相色谱仪[配四元梯度泵(G1311C 型)、自动进样器(G1329B 型)、柱温箱(G1316A 型)、DAD 检测器(G4212B 型)、ChemStation 工作站]、Agilent ZORBAX Plus-C_{18} 色谱柱(250 mm×4.6 mm，5 μm)(美国 Agilent 公司)；PS-100A 超声波清洗器(东莞洁康超声波设备有限公司)；XP204 型万分之一电子天平(瑞士梅特勒-托利多公司)；GL21M 快速冷冻离心机(上海卢湘仪离心机仪器有限公司)；QL-901 型旋涡混合器(海门其林贝尔仪器制造有限公司)；XL-08B 型摇摆式粉碎机(广州市旭朗机械设备有限公司)；0.22 μm 微孔滤膜(天津津腾公司)。

1.3 方 法

1.3.1 对照品和样品溶液制备

分别精密称取没食子酸、表没食子儿茶素、表儿茶素没食子酸酯、表儿茶素、儿茶素、咖啡碱和表没食子儿茶素没食子酸酯对照品各 20 mg，加入甲醇经超声溶解后，即得 10 mg/mL 质量浓度的对照品溶液。

精密称取样品黑茶细粉(经粉碎过 40 目筛)0.20 g，置于 50 mL 容量瓶中，加入甲醇进行溶解，水浴(60°C)中超声提取 60 min，静置至室温，取上清液以 13000 r/min 转速离心 10 min 后，经 0.22 μm 滤膜滤过，取滤液即得样品溶液。

1.3.2 色谱条件

流动相 A: 5% (*m*/*V*)乙酸水；流动相 B: 乙腈；柱温 35°C；检测波长: 280 nm；进样量: 20 μL；流速: 0.3 mL/min；梯度洗脱程序如表 2 所示。

表 2 梯度洗脱程序

Table 2 Gradient elution program

时间/min	A/%	B/%
0.0	85	15
8.0	85	15
25.0	75	25
35.0	60	40
45.0	40	60
55.0	20	80
60.0	10	90

1.3.3 指纹图谱方法学考察

按照 1.3.1 所述方法制备 1 份样品溶液并在 1.3.2 条件下连续进样 6 次，计算 HPLC 图谱中各共有峰相对保留时

间和相对峰面积的相对标准偏差(relative standard deviations, RSDs)用于评价精密度。按照 1.3.1 所述方法平行制备 6 份样品溶液并在 1.3.2 条件下分别进样, 计算 HPLC 图谱中各共有峰相对保留时间和相对峰面积的 RSDs 值用于评价重复性。取同一样品溶液在不同时间点(0、4、8、12、16、20、24 h)在 1.3.2 条件下连续进样 6 次, 计算 HPLC 图谱中各共有峰的相对保留时间和相对峰面积的 RSDs 值用于评价稳定性。

1.3.4　不同产地黑茶 HPLC-DAD 指纹图谱的建立及共有峰指认

(1)指纹图谱建立

取 48 份黑茶按照 1.3.1 所述方法制备样品溶液, 在 1.3.2 条件下进行检测, 所获得的色谱数据以 AIA 格式导入《中药色谱指纹图谱相似度评价系统》(2012 版)中, 设置 S2 号黑茶图谱作为参照峰, 时间窗宽度 0.5 min, 选择多点校正方法进行色谱峰匹配, 生成 5 个产地黑茶指纹图谱和共有模式。

(2)共有峰指认

按照 1.3.1 所述方法制备的 7 种对照品溶在 1.3.2 条件下液连续进样 6 次, 计算同源峰保留时间的滞后值, 考察 5 个不同产地黑茶的 HPLC-DAD 指纹图谱共有模式中各共有峰的平均保留时间, 根据同源峰保留时间的滞后值对不同产地黑茶的共有峰进行指认。

1.4　产地识别模型构建方法和评价指标

1.4.1　产地识别模型构建方法

PLS-DA 是一种基于偏小二乘回归的线性判别方法[29]。应用 PLS-DA 模型进行分类时, 通过在自变量矩阵(X)和因变量矩阵(Y)之间建立偏最小二乘回归模型, Y 因变量矩阵使用二进制编码(1 或 0)来表示, 其中 1 表示样本属于该类, 0 表示不属于该类[34−35]。LR 是一种广义线性回归分析模型[36], 通过建立代价函数不断优化方法迭代求解最优模型参数, 用于解决二分类问题[37]。BPNN 是由非线性变换神经单元组成的一种前馈多层神经网络, 一般包括: 一个输入层、一个或多个隐层和一个输出层[38], 其基本思想是通过传播的前向和反向共同作用, 不断调整各层的权值和阈值, 从而获取误差最小的预测值[39]。SVM 是一种二分类器, 该算法能够解决非线性和高维问题, 通过将低维线性不可区分的原始数据映射到高维的特征空间, 在特征空间上建立一个超平面作为决策曲面[40], 使其线性可分并具有良好的泛化能力[41], 并且在小样本集中有效性较高[42]。RF 是一种有监督的非线性算法[43], 其基本思想是通过集成多个决策树分类器投票或取均值方式而获得预测分类结果, 该方法其识别准确率比单个分类器高且无需对高维数据进行降维。DT 是一种非参数监督学习方法, 其模型生成过程是一个递归过程[44]。该算法基本思想是通过多个树节点的判别获取已知样本类别信息, 从而提炼出树型分类模型, 树中的根结点到每个叶结点对应了一个判定测试序列, 以此预测未知样本的类别。

1.4.2　评价指标

采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)及 F_1 分数(F_1 Score)作为黑茶产地识别模型的评价指标。Accuracy 是评估黑茶产地识别模型的预测准确度的指标, 越高说明模型的产地预测正确率越好; Precision 表示 48 份黑茶中预测为正确产地的数量所占比重, 越高说明模型的特异鉴别能力越好; Recall 代表了黑茶产地识别模型正确预测的比率, 越高说明模型对各产地黑茶的正确识别能力越好; F_1 Score 是综合考虑精确率和召回率的调和平均数, 最大值为 1, 最小值为 0。其计算式分别为:

$$\text{Accuracy}=(TP+TN)/(TP+FP+FN+TN)\times 100\% \quad (1)$$

$$\text{Precision}=(TP)/(TP+FP)\times 100\% \quad (2)$$

$$\text{Recall}=(TP)/(TP+FN)\times 100\% \quad (3)$$

$$F_1\ \text{Score}=(2\times \text{Precision}\times \text{Recall})/(\text{Precision}+\text{Recall}) \quad (4)$$

其中, TP 为被预测正确的该产地黑茶样本的数量; TN 为被预测正确的其他产地黑茶样本的数量; FP 为被错误预测为该产地黑茶样本的数量; FN 为被错误预测为其他产地黑茶样本的数量。

1.5　处理软件

采用中药指纹图谱相似度评价系统(2012 版)建立 HPLC 指纹图谱, 采用 Origin Pro 2023b 软件绘制图谱, 采用 Simca 14.1 软件和 Python 软件利用指纹图谱中的共有峰面积分别建立 PLS-DA、LR、BPNN、SVM、RT 和 DT 产地识别模型。

2　结果与分析

2.1　指纹图谱方法学验证

指纹图谱方法学验证结果显示, 在黑茶 HPLC 的精密度实验中, 各共有峰的相对保留时间 RSDs 值为 0.03%~0.78%, 相对峰面积 RSDs 值为 0.86%~2.61%, 均小于 3% (n=6), 表明实验仪器精密度良好; 在重复性实验中, 计算得到各共有峰的相对保留时间 RSDs 值为 0.03%~0.62%, 相对峰面积 RSDs 值为 0.87%~2.63%, 均小于 3% (n=6), 表明该实验方法重复性良好; 在稳定性实验中, 计算得到各共有峰相对保留时间 RSDs 值为 0.01%~0.24%, 相对峰面积 RSDs 值为 0.76%~2.46%, 均小于 3% (n=6), 表明样品溶液在 24 h 内稳定性良好。

2.2　HPLC 指纹图谱建立及共有峰指认

采用中药色谱指纹图谱相似度评价系统对来自我国 5 个不同产地的 48 份黑茶进行多点矫正后所得的 HPLC-DAD 指纹图谱见图 1, 共标记 8 个共有峰, 总峰面积占 90%以上, 进一步将 48 份黑茶指纹图谱共有模式(图 2b)与标准品图谱(图 2a)进行比较后, 分别确定了 1 号峰为

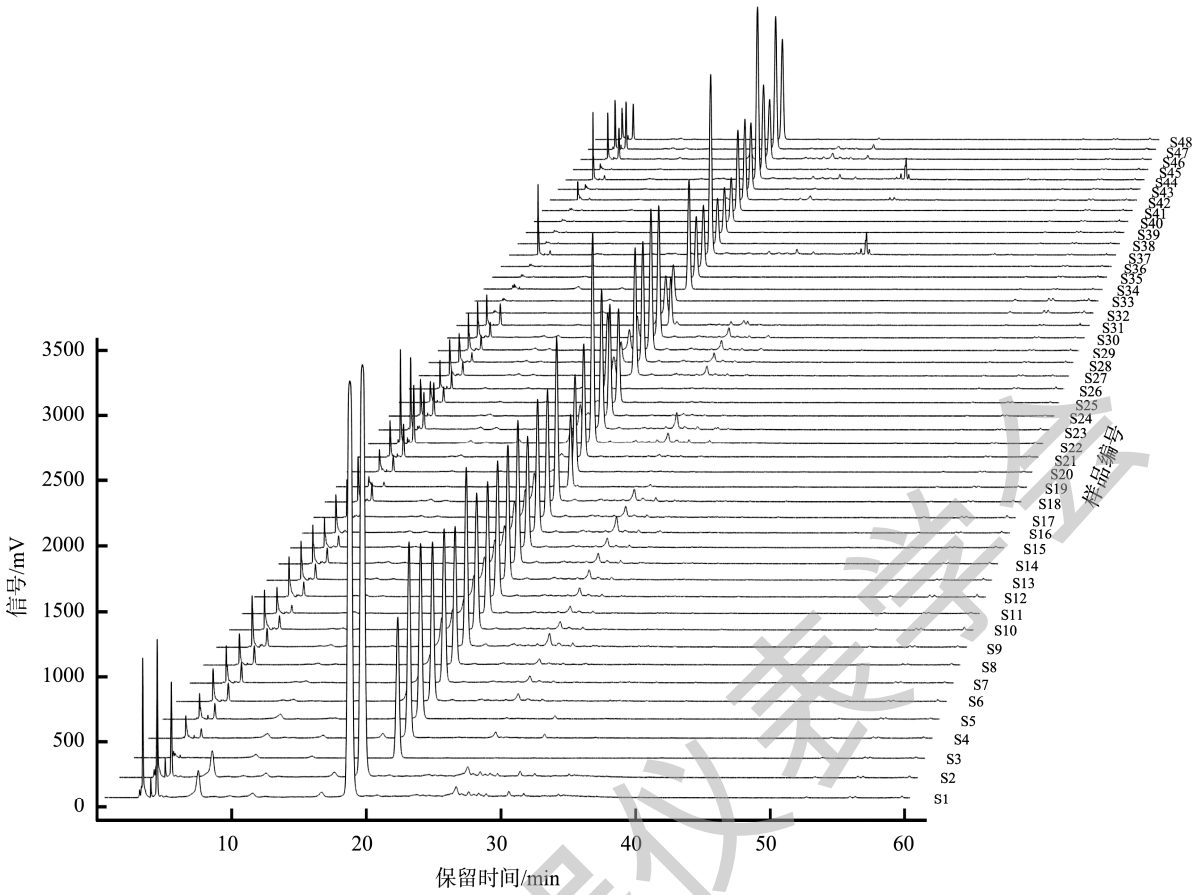


图 1 48 份黑茶 HPLC-DAD 的指纹图谱

Fig.1 HPLC-DAD fingerprint of 48 dark tea samples

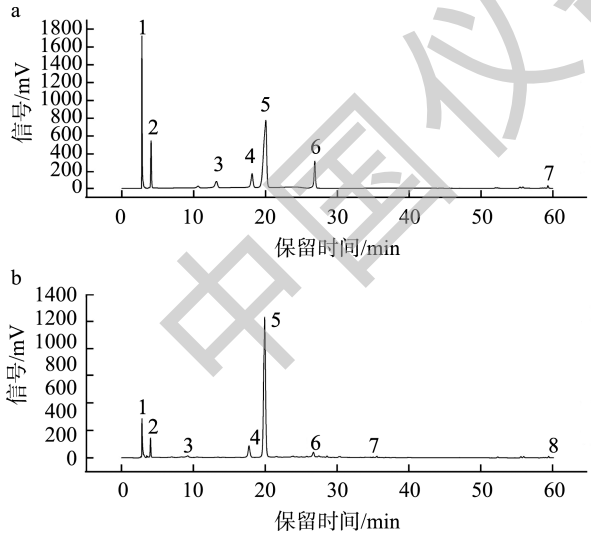


图 2 黑茶指纹图谱标品对照图(a)和共有模式图(b)

Fig.2 Dark tea fingerprint standard control chart (a) and common pattern chart (b)

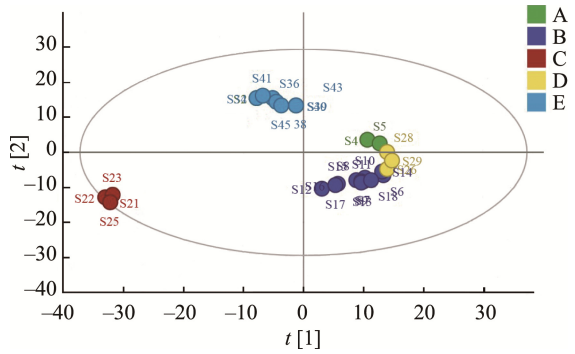
为没食子酸(t_R=2.800 min)、2 号峰为表没食子儿茶素(t_R=4.086 min)、3 号峰为儿茶素(t_R=13.168 min)、4 号峰为表儿茶素(t_R=18.141 min)、5 号峰为咖啡碱(t_R= 20.054 min)、6 号峰为表没食子儿茶素没食子酸酯(t_R=26.781 min)、8 号峰为表儿茶素没食子酸酯(t_R=59.341 min), 7 号峰未确定其成分。以 S2 号黑茶图谱作为参照峰, 计算 8 个共有峰相对保留时间的 RSDs 为 0.03%~0.38%, 相差较小; 相对峰面积的 RSDs 为 31.55%~80.61%, 相差较大。由此可知, 不同产地黑茶化学成分种类上具有一致性, 但同一成分的含量差异明显, 成分含量差异可作为区分不同产地黑茶的主要依据, 用于黑茶产地识别模型的学习与建立。

2.3 黑茶产地识别模型的建立与分析

2.3.1 基于 PLS-DA 黑茶产地识别模型的建立

本研究利用 8 个共有峰的峰面积作为特征向量对不同产地黑茶进行产地识别, 在建立黑茶的产地识别模型前, 使用数据标准化与自动缩放步骤对原始数据进行了预处理, 增加数据可靠性[45], 使数据质量得到提升, 并以不同产地作为因变量(Y)建立了 5 个产地黑茶的 PLS-DA 产地识别模型。由 PLS-DA 的得分图(图 3)可知, 不同地区的黑茶明显分为 5 类, 均无超过置信区间 95%的点, 无异常值存在。模型自变量拟合指数(R^2_X)和因变量拟合指数(R^2_Y)分别为 0.978 和 0.708, 模型预测指数(Q^2)为 0.542, R^2 和 Q^2 均超过 0.500, 说明模型有较好的稳定性和预测能力[46]。

通过 PLS-DA 得到 8 个特征变量的重要性投影(variable importance in projection, VIP)结果如图 4 所示。以 VIP>1 作为显著影响的衡量指标，共找到 4 个差异标志物，这几个化学成分对于区分不同产地黑茶的贡献较大，推测这 4 个化学成分可为不同产地黑茶的差异性成分，对其影响显著性从大到小排序，分别为表儿茶素没食子酸酯>儿茶素>没食子酸>表儿茶素。图 5 为 200 次置换检验结果图，图中 Q^2 回归线与纵轴的相交点小于零，说明模型不存在过拟合现象，模型验证有效[47]，可认为该结果可用于黑茶的产地识别分析。



注：A、B、C、D、E 分别为来自广西、湖南、四川、陕西和西藏的黑茶样品。

图 3　PLS-DA 模型分类得分图

Fig.3　PLS-DA model classification score chart

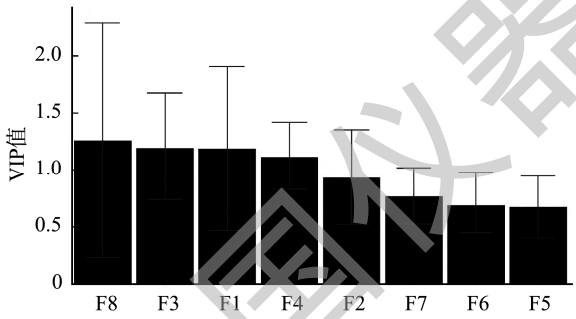


图 4　PLS-DA 模型 VIP 值图

Fig.4　PLS-DA model VIP value chart

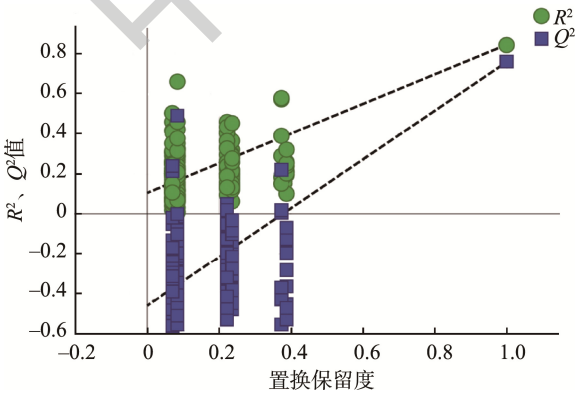


图 5　PLS-DA 的 200 次置换检验图

Fig.5　200-time permutation test chart of PLS-DA

2.3.2　基于不同机器学习算法的黑茶产地识别模型建立

PLS-DA 得分图(图 3)显示，来自广西与西藏和湖南与陕西两两地区间的黑茶分类存在重叠情况，为进一步探究多元统计分析在黑茶产地识别的准确性和可靠性，本研究采用 Zscore 标准化对 8 个共有峰峰面积组成的原始数据阵列进行预处理，利用 LR、BPNN、SVM、RT 和 DT 学习算法建立黑茶产地识别模型，并将数据集划分为测试集和预测集，最终预测分类结果如图 6 和表 3 所示。SVM 模型的 Precision 指标结果虽然低于 RT 和 DT 模型，但是 Accuracy、Recall 和 F_1 Score 指标结果最优，分别为 0.900、0.900、0.899。因此从整体上分析，SVM 相较于其他 4 种模型的产地识别效果最好[37]。

进一步对结果分析发现，BPNN 和 SVM 的预测准确率均达到了 90%，表明这两个机器学习模型较适合黑茶产地识别研究；RT 和 DT 的预测准确率虽然低于 BPNN，但 Precision 和 F_1 Score 指标均优于 BPNN，其中两者 Precision 为 1.000，说明 RT 和 DT 具有良好特异鉴别能力；相比 BPNN、SVM、RT 和 DT 模型的所有模型评价指标，LR 模型的预测识别效果最差。

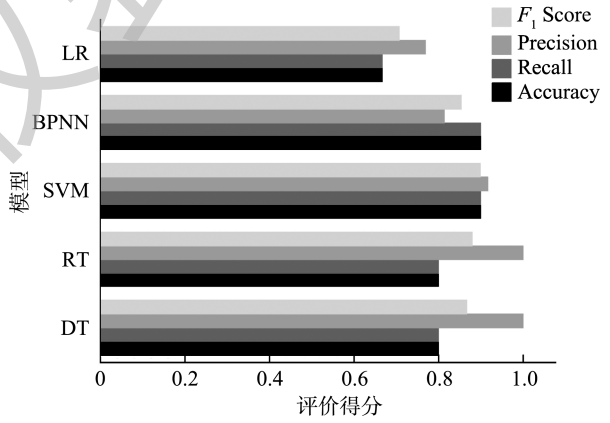


图 6　不同产地黑茶分类效果图

Fig.6　Classification effects of dark tea from different origins

表 3　不同产地黑茶预测分类结果

Table 3　Predicted classification results of dark tea from different origins

模型	预测集			
	F_1 Score	Recall	Precision	Accuracy
LR	0.707	0.667	0.769	0.667
BPNN	0.854	0.900	0.814	0.900
SVM	0.899	0.900	0.917	0.900
RT	0.880	0.800	1.000	0.800
DT	0.867	0.800	1.000	0.800

3　结　论

本研究利用 HPLC-DAD 建立我国不同产地黑茶的指

纹图谱，确定 8 个共有峰并指认其中 7 种成分。通过计算 8 个共有峰的相对保留时间和相对峰面积的 RSDs 值发现，不同产地黑茶化学成分种类上具有一致性，但同一成分的含量存在差异，这与黑茶生产中原料品种、采摘期、发酵程度、拼配比例、贮存条件等因素有关。此外，结合化学计量学和机器学习模型对黑茶产地进行溯源研究发现，基于 PLS-DA 化学计量学模型共找到 4 个差异标志物，其影响显著性从大到小依次为表儿茶素没食子酸酯>儿茶素>没食子酸>表儿茶素，模型预测准确率为 54.2%。采用 5 种机器学习模型进行黑茶产地识别的结果显示，其中 BPNN、SVM、RT 和 DT 产地识别模型对来自 5 个产地黑茶的预测准确率均达到或高于 80.0%，并获得较好预测分类精度；结合评价指标对比分析后表明，其中 SVM 模型在识别黑茶产地整体效果上及性能最佳；进一步与 PLS-DA 的预测准确率比较后发现，5 种机器学习模型的产地识别效果均优于 PLS-DA，表明 HPLC 指纹图谱结合机器学习算法能够较好对黑茶产地进行溯源研究。

综上，本研究采用 HPLC-DAD 结合机器学习进行黑茶产地识别研究，并验证该方法在小样本集中实现黑茶产地溯源具有可行性和有效性，这为我国黑茶产地溯源研究提供参考。但是本研究收集的黑茶仅来自我国陕西、四川、湖南、广西和西藏产地共 48 个黑茶样品，部分地区样品数量有限，无法代表我国黑茶样品整体情况，后续可进一步增加样品量用于黑茶产地识别模型的训练，以获得更佳的预测识别效果和模型性能。

参考文献

[1] 胡燕, 齐桂年. 四川黑茶的高效液相色谱指纹图谱研究[J]. 西北农林科技大学学报(自然科学版), 2015, 43(1): 134–140.
HU Y, QI GN. High performance liquid chromatographic fingerprinting of Sichuan dark tea [J]. J Northwest Agric Forest Univ (Nat Sci Ed), 2015, 43(1): 134–140.

[2] 王冰洁, 黄奕, 赵芸, 等. 黑茶品质化学成分及工艺影响因素研究进展[J]. 食品工业, 2023, 44(1): 206–211.
WANG BJ, HUANG Y, ZHAO Y, *et al*. Research progress on chemical composition and process influencing factors of dark tea quality [J]. Food Ind, 2023, 44(1): 206–211.

[3] 党旭辉, 周秦羽, 刘梦圆, 等. 黑茶金花体外抗氧化及降血脂活性研究[J]. 食品安全质量检测学报, 2022, 13(24): 7927–7933.
DANG XH, ZHOU QY, LIU MY, *et al*. Study on antioxidant and hypolipidemic activity *in vitro* of post-flowering dark tea [J]. J Food Saf Qual, 2022, 13(24): 7927–7933.

[4] ZHU J, ZHOU H, ZHANG J, *et al*. Valorization of polysaccharides obtained from dark tea: Preparation, physicochemical, antioxidant, and hypoglycemic properties [J]. Foods, 2021, 10(10): 2276.

[5] 苏超, 时庆欣, 黄荣增, 等. 黑茶对高脂血症小鼠的降血脂作用及其药效物质基础研究[J]. 湖北中医药大学学报, 2022, 24(3): 42–45.
SU C, SHI QX, HUANG RZ, *et al*. Preventive effect of dark tea on high fat diet induced hyperlipidemic mice and its bioactive components [J]. J Hubei Univ Chin Med, 2022, 24(3): 42–45.

[6] MA W, SHI Y, YANG G, *et al*. Hypolipidaemic and antioxidant effects of various Chinese dark tea extracts obtained from the same raw material and their main chemical components [J]. Food Chem, 2022, 375: 131877.

[7] 汤荃荃, 湛莉, 张梓莹, 等. 黑茶桑叶固体饮料对高脂饮食小鼠的减肥作用[J]. 中国酿造, 2022, 41(6): 164–170.
TANG QQ, ZHAN L, ZHANG ZY, *et al*. Effect of dark tea and mulberry leave solid beverage on weight loss of high fat diet mice [J]. China Brew, 2022, 41(6): 164–170.

[8] QU J, YE M, WEN C, *et al*. Compound dark tea ameliorates obesity and hepatic steatosis and modulates the gut microbiota in mice [J]. Front Nutr, 2023, 10: 1082250.

[9] 傅冬和, 朱洛志, 刘仲华. 安化黑茶调理肠胃功效及作用机理[J]. 中国茶叶, 2023, 45(3): 1–7.
FU DH, ZHU MZ, LIU ZH. Effect and mechanism of Anhua dark tea on regulating gastrointestinal tract [J]. China Tea, 2023, 45(3): 1–7.

[10] GONG Z, OUYANG J, WU X, *et al*. Dark tea extracts: Chemical constituents and modulatory effect on gastrointestinal function [J]. Biomed Pharmacother, 2020, 130: 110514.

[11] 王叶, 孟涛, 戴学文, 等. “安化黑茶”地理标志保护现状及发展对策探究[J]. 福建茶叶, 2023, 45(1): 136–139.
WANG Y, MENG T, DAI XW, *et al*. Anhua dark tea’ geographical indication protection status and development countermeasures [J]. Tea Fujian, 2023, 45(1): 136–139.

[12] 黄静, 林海, 何畅辉, 等. 基于茶叶感官审评技术浅述各产地黑茶品质特征[J]. 湖北农业科学, 2020, 59(S1): 396–399.
HUANG J, LIN H, HE CH, *et al*. Analyses on the characters in producting areas of dark tea based on sensory evaluation [J]. Hubei Agric Sci, 2020, 59(S1): 396–399.

[13] 颜鸿飞, 彭争光, 李蓉娟, 等. GC-TOF MS 结合化学计量学用于安化黑茶的识别[J]. 食品与机械, 2017, 33(8): 34–37, 65.
YAN HF, PENG ZG, LI RJ, *et al*. Discrimination of Anhua dark tea by gas chromatography-time of flight mass spectrometry combined with chemometrics [J]. Food Mach, 2017, 33(8): 34–37, 65.

[14] 胡丽珍. 基于 LC-MS 的不同茶类特征化学变量筛选及判别方法建立[D]. 合肥: 安徽农业大学, 2020.
HU LZ. Selection of characteristic chemical variables and establishment of discriminant method for different kinds of tea based on LC-MS [D]. Hefei: Anhui Agricultural University, 2020.

[15] LV S, WU Y, ZHOU J, *et al*. The study of fingerprint characteristics of Dayi Pu-erh tea using a fully automatic HS-SPME/GC–MS and combined chemometrics method [J]. PLoS One, 2014, 9(12): e116428.

[16] SHUAI M, PENG C, NIU H, *et al*. Recent techniques for the authentication of the geographical origin of tea leaves from *Camellia sinensis*: A review [J]. Food Chem, 2022, 374: 131713.

[17] LIU H, ZENG Y, ZHAO X, *et al*. Improved geographical origin discrimination for tea using ICP-MS and ICP-OES techniques in

combination with chemometric approach [J]. J Sci Food Agric, 2020, 100(8): 3507–3516.

[18] 袁园, 唐延林. 偏最小二乘法结合主成分分析对黑茶产地的研究[J]. 大学物理实验, 2020, 33(1): 50–55.
YUAN Y, TANG YL. A study of dark tea origins by partial least squares combined with principal component analysis [J]. Phys Exp Coll, 2020, 33(1): 50–55.

[19] 张婉, 胡文文, 叶巧茹, 等. 茶叶产地溯源研究进展[J]. 亚热带农业研究, 2021, 17(2): 137–144.
ZHANG W, HU WW, YE QR, *et al*. Research progress on geographical origin of tea (*Camellia sinensis*) [J]. Subtropical Agric Res, 2021, 17(2): 137–144.

[20] 陈波, 张巍, 康海宁, 等. 茶叶的 ^{1}H NMR 指纹图谱研究[J]. 波谱学杂志, 2006, (2): 169–180.
CHEN B, ZHANG W, KANG HN, *et al*. ^{1}H NMR fingerprinting study of tea leaves [J]. Chin J Magn Reson, 2006, (2): 169–180.

[21] 钟妮, 赵熙, 黄浩, 等. 黑茶产地溯源技术研究进展[J]. 茶叶通讯, 2023, 50(1): 24–31.
ZHONG N, ZHAO X, HUANG H, *et al*. Advances in origin traceability technology of dark tea [J]. J Tea Commun, 2023, 50(1): 24–31.

[22] YUDTHAVORASIT S, WONGRAVEE K, LEEPIPATPIBOON N. Characteristic fingerprint based on gingerol derivative analysis for discrimination of ginger (*Zingiber officinale*) according to geographical origin using HPLC-DAD combined with chemometrics [J]. Food Chem, 2014, 158: 101–111.

[23] WU Q, GU HW, YIN XL, *et al*. Development of an HPLC-DAD method combined with chemometrics for differentiating geographical origins of Chinese red wines on the basis of phenolic compounds [J]. Food Anal Method, 2021, 14: 1895–1907.

[24] REN H, YUE JP, WANG DD, *et al*. HPLC and H-1-NMR combined with chemometrics analysis for rapid discrimination of floral origin of honey [J]. J Food Meas Charact, 2019, 13: 1195–1204.

[25] 胡燕. 不同产地黑茶的化学指纹图谱研究[D]. 成都: 四川农业大学, 2014.
HU Y. Study on chemical fingerprinting of dark teas from different regions [D]. Chengdu: Sichuan Agricultural University, 2014.

[26] 常睿, 马梦君, 罗理勇, 等. 基于生化成分构建不同地区黑茶分类模型[J]. 食品与发酵工业, 2019, 45(11): 91–98.
CHANG R, MA MJ, LUO LY, *et al*. The classification model of dark tea in different regions was constructed based on biochemical components [J]. Food Ferment Ind, 2019, 45(11): 91–98.

[27] 白秀芝, 王美玲, 颜鸿飞, 等. 高效液相色谱指纹图谱及随机森林应用于湖南安化黑茶水溶性成分的研究[J]. 分析测试学报, 2014, 33(11): 1268–1273.
BAI XZ, WANG ML, YAN HF, *et al*. Investigation on fingerprint of hunan anhua dark tea's water-soluble components by high performance liquid chromatography and random forest [J]. J Instrum Anal, 2014, 33(11): 1268–1273.

[28] 刘静, 亓超凡, 绪扩, 等. 指纹图谱技术在食品质量与安全中的应用研究进展[J]. 食品安全质量检测学报, 2022, 13(10): 3189–3197.
LIU J, QI CF, XU K, *et al*. Research progress on the application of fingerprint technology in the field of food quality and safety [J]. J Food Saf Qual, 2022, 13(10): 3189–3197.

[29] CUI C, XIA M, WEI Z, *et al*. 1H NMR-based metabolomic approach combined with machine learning algorithm to distinguish the geographic origin of huajiao (*Zanthoxylum bungeanum* Maxim) [J]. Food Control, 2023, 145: 109476.

[30] YUN J, CUI CJ, ZHANG SH, *et al*. Use of headspace GC/MS combined with chemometric analysis to identify the geographic origins of black tea [J]. Food Chem, 2021, 360: 130033.

[31] JIN G, XU YF, CUI CJ, *et al*. Rapid identification of the geographic origin of Taiping Houkui green tea using near-infrared spectroscopy combined with a variable selection method [J]. J Sci Food Agric, 2022, 102: 6123–6130.

[32] HU Y, XU LJ, HUANG P, *et al*. Reliable identification of Oolong tea species: Nondestructive testing classification based on fluorescence hyperspectral technology and machine learning [J]. Agriculture-Basel, 2021, 11(11): 1106.

[33] NING JM, SUN JJ, LI SH, *et al*. Classification of five Chinese tea categories with different fermentation degrees using visible and near-infrared hyperspectral imaging [J]. Int J Food Prop, 2017, 20: 1515–1522.

[34] SACCENTI E, TIMMERMAN ME. Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data [J]. J Proteome Res, 2016, 15(8): 2379–2393.

[35] ZONTOV YV, RODIONOVA OY, KUCHERYAVSKIY SV, *et al*. PLS-DA–A MATLAB GUI tool for hard and soft approaches to partial least squares discriminant analysis [J]. Chemom Intell Lab Syst, 2020, 203: 104064.

[36] YE WX, YAN TY, ZHANG C, *et al*. Detection of pesticide residue level in grape using hyperspectral imaging with machine learning [J]. Foods, 2022, 11(11): 1609.

[37] 王劭晟, 田绪红, 邱少健, 等. 基于近红外光谱结合机器学习的鳕鱼品种二分类方法研究[J]. 食品安全质量检测学报, 2021, 12(22): 8651–8659.
WANG SS, TIAN XH, QIU SJ, *et al*. Research on cod species binary classification method based on near infrared spectroscopy and machine learning [J]. J Food Saf Qual, 2021, 12(22): 8651–8659.

[38] YANG ZW, WANG ZQ, YUAN WH, *et al*. Classification of wolfberr from different geographical origins by using electronic tongue and deep learning algorithm [C]. International-Federation-of-Automatic-Control (IFAC) Conference on Sensing, Control and Automation Technologies for Agriculture (AGRICONTROL), 2019.

[39] PENG KD, SHANG YZ, KE XZ, *et al*. Rapid identification of ginseng origin by laser induced breakdown spectroscopy combined with neural network and support vector machine algorithm [J]. Acta Phys Sin, 2021, 70(4): 040201.

[40] 王靖会, 臧妍宇, 曹崴, 等. 基于机器学习方法的吉林大米产地确证模

型研究[J]. 中国粮油学报, 2018, 33(9): 123–130.
WANG JH, ZANG YY, CAO W, *et al*. Research on Jilin rice origin confirmation model based on machine learning methods [J]. J Chin Cereals Oils Assoc, 2018, 33(9): 123–130.

[41] CHEN Q, ZHAO J, FANG CH, *et al*. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM) [J]. Spectrochim Acta A, 2007, 66(3): 568–574.

[42] 罗雅文, 董大明. 激光诱导击穿光谱结合化学计量学的淫羊藿产地快速鉴别[J]. 西北师范大学学报(自然科学版), 2023, 59(2): 54–60, 71.
LUO YW, DONG DM. Rapid identification of the origin of *Epimedium* by laser-induced breakdown spectroscopy combined with chemometrics [J]. J Northwest Norm Univ (Nat Sci), 2023, 59(2): 54–60, 71.

[43] 龚圣, 朱雅宁, 曾陈娟, 等. 近红外光谱结合随机森林算法:一种快速有效的附子产地溯源策略[J]. 光谱学与光谱分析, 2022, 42(12): 3823–3829.
GONG S, ZHU YY, ZENG CJ, *et al*. Near-infrared spectroscopy combined with random forest algorithm: A fast and effective strategy for origin traceability of Fuzi [J]. Spectrosc Spect Anal, 2022, 42(12): 3823–3829.

[44] WANG HX, CUI WJ, GUO YC, *et al*. Machine learning prediction of foodborne disease pathogens: Algorithm development and validation study [J]. Jmir Med Inform, 2021, 9(1): e24924.

[45] 金承亮, 王永军, 黄河, 等. 高维红外光谱数据预处理在中药材产地鉴别中的应用[J]. 光谱学与光谱分析, 2023, 43(7): 2238–2245.
JIN CL, WANG YJ, HUANG H, *et al*. Application of high-dimensional infrared spectral data preprocessing inthe origin identification of traditional Chinese medicinal materials [J]. Spectrosc Spect Anal, 2023, 43(7): 2238–2245.

[46] 范雪花, 王艳丽, 侯富国, 等. 基于 HPLC 指纹图谱结合化学计量学的白及与黄花白及辨识研究[J]. 中草药, 2023, 54(12): 3990–3998.
FAN XH, WANG YL, HOU FG, *et al*. Identification of *Bletilla striata* and *Bletilla ochracea* by HPLC coupled with chemometrics [J]. Chin Tradit Herbal Drugs, 2023, 54(12): 3990–3998.

[47] 赵永恒, 张勇, 秦志旺, 等. 气相色谱-质谱指纹图谱结合化学计量学方法分析不同产地艾叶挥发油的差异[J]. 理化检验-化学分册, 2022, 58: 1277–1282.
ZHAO YH, ZHANG Y, QIN ZW, *et al*. Analysis of differences of volatile oil in *Folium artemisiae* Argyi from different origins by GC-MS fingerprint chromatograms combined with chemometrics [J]. Phys Test Chem Anal Part B, 2022, 58: 1277–1282.

(责任编辑: 郑 丽 张晓寒)

作者简介

王贞红, 硕士, 副教授, 主要研究方向为茶树栽培和茶叶加工。
E-mail: wzhxzlz@xza.edu.cn

韩沅汐, 硕士研究生, 主要研究方向为高原特色农产品产地溯源研究。
E-mail: 903767780@qq.com

李 梁, 硕士, 副教授, 主要研究方向为高原特色农产品功能成分挖掘与产地溯源。
E-mail: jwllok@sina.com