物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

激光诱导击穿光谱技术结合神经网络和支持向量机算法的人参产地快速识别研究

董鹏凯 赵上勇 郑柯鑫 王蓟 高勋 郝作强 林景全

Rapid identification of ginseng origin by laser induced breakdown spectroscopy combined with neural network and support vector machine algorithm

Dong Peng-Kai Zhao Shang-Yong Zheng Ke-Xin Wang Ji Gao Xun Hao Zuo-Qiang Lin Jing-Quan

引用信息 Citation: Acta Physica Sinica, 70, 040201 (2021) DOI: 10.7498/aps.70.20201520
在线阅读 View online: https://doi.org/10.7498/aps.70.20201520
当期内容 View table of contents: http://wulixb.iphy.ac.cn

激光诱导击穿光谱技术结合神经网络和支持向量机算法的人参产地快速识别研究*

董鹏凯[1)] 赵上勇[1)] 郑柯鑫[1)] 王蔺[1)†] 高勋[1)‡] 郝作强[2)] 林景全[1)]

1) (长春理工大学理学院, 长春 130022)

2) (山东师范大学物理与电子科学学院, 济南 250358)

(2020 年 9 月 11 日收到; 2020 年 10 月 19 日收到修改稿)

利用激光诱导击穿光谱技术结合机器学习算法, 对东北 5 个产地 (大兴安岭、集安、恒仁、石柱、抚松) 的人参进行产地识别, 建立了主成分分析算法分别结合反向传播 (BP) 神经网络和支持向量机算法的人参产地识别模型. 实验采集了 5 个产地人参共 657 组在 200—975 nm 的激光诱导击穿光谱, 经光谱数据预处理后, 对 C, Mg, Ca, Fe, H, N, O 等元素的 8 条特征谱线进行主成分分析, 原光谱数据的前 3 个主成分累积贡献率达到 92.50%, 且样品在主成分空间中呈现良好的聚集分类. 降维后的前 3 个主成分以 2∶1 进行随机抽取, 分别作为分类算法的训练集和测试集. 实验结果表明主成分分析结合 BP 神经网络及支持向量机的平均识别率分别为 99.08% 和 99.5%. 发生误判的原因是集安和石柱两地地理环境的接近而导致的 H, O 两元素在 Ca 元素离子发射谱线下的归一化强度相似. 本研究为激光诱导击穿光谱技术在人参产地的快速识别提供了方法和参考.

关键词: 激光诱导击穿光谱, 机器学习算法, 产地识别, 人参

PACS: 02.10.Yn, 33.15.Vb, 98.52.Cf, 78.47.dc **DOI:** 10.7498/aps.70.20201520

1 引 言

人参 (panax ginseng) 是五加科多年生草本植物, 在中国已有 4000 多年的药用和食用历史. 人参中主要有效成分为人参皂苷和多糖, 还含有维生素类、酶类、有机酸及其酯、蛋白质、甾醇及其苷、多肽类、含氮化合物、木质素、黄酮类和无机元素等多种成分, 具有滋补强身、预防疲劳、抗衰老、抗肿瘤、提高免疫功能等多种功效, 被广泛应用于制药、保健产品、美容产品、饮料等领域, 对内分泌系统、心血管疾病和中枢神经系统等方面有突出疗效[1,2]. 研究发现, 人参皂苷、多糖等主要有效成分在人参内形成、转化与积累等过程与人参产地的土壤环境、日照环境和气候环境有关, 因此不同人参产地的相同品种人参在临床疗效上存在着较大的差异. 目前, 中国人参产地众多, 同一品种人参质量参差不齐, 质量监控困难. 东北三省是我国重要的人参产地, 目前不法商人借“长白山人参”等噱头出售人参来牟取利益, 导致人参市场充斥大量伪品及混淆品, 严重影响人参的有效使用以及国际市场的推广. 所以人参产地的识别对人参质量品牌保护非常重要, 并且对提高中药制剂的临床疗效均一性和稳定性及人参市场的发展具有重要研究意义.

* 国家自然科学基金 (批准号: 61575030)、吉林省自然科学基金 (批准号: 20180101283JC, 20200301042RQ, 20180201033GX, 20190302125GGX) 和吉林省教育厅 (批准号: JJKH20190539KJ) 资助的课题.

† 通信作者. E-mail: jiji_w@163.com

‡ 通信作者. E-mail: gaoxun@cust.edu.cn

传统的“五行”“六体”识别方法对人参种类和质量的判断易受人为因素影响. 随着现代科技的发展, 通过对药效成分含量的测定来确定不同产地药材的差异是重要的中草药识别方法. 光谱技术因能客观地反映药材内在质量从而被广泛应用于中草药鉴定中, 常用的光谱检测方法主要有近红外光谱 (near infrared spectroscopy, NIR) 技术、拉曼光谱 (Raman spectroscopy) 技术、荧光光谱 (fluorescence spectroscopy) 技术等[3−6]. 常规的光谱技术由于光谱信号微弱很容易受到背景光的影响, 且检测样品时处理时间长且复杂, 无法实现实时、在线和快速检测. 因此, 亟需一种快速可靠的人参产地检测方法.

激光诱导击穿光谱技术 (laser inducted breakdown spectroscopy, LIBS) 是一种原子发射光谱技术[7−9], 适用于所有物质 (气态、液态、固态), 具有快速、微损、样品准备简单和多元素同时探测等优点, 广泛地应用于爆炸物检测[10]、文化遗产[11]、生物医学分析[12]、土壤重金属检测[13]、地质分析[14]、食品安全[15] 等领域. 利用 LIBS 技术和化学计量学方法结合可实现待测样品的分类识别. Junjuri 和 Gundawar[16] 利用主成分分析 (principal component analysis, PCA) 方法和人工神经网络 (artificial neural network, ANN) 两种算法结合 LIBS 技术, 采用 PCA 方法对样品进行分析, 以主成分数据作为 ANN 的输入量实现了对 5 种消费塑料进行鉴定, 最终识别精确度为 97%—99%; Velioglu 等[17] 利用 LIBS 结合 PCA 实现了纯下脚料和混合下脚料掺假牛肉样品的识别; Lin 等[18] 使用 LIBS 技术结合偏最小二乘 (PLS-LDA) 及支持向量机 (support vector machines, SVM) 方法实现了钢种的识别, 采用偏最小二乘支持向量机算法 (LSSVM) 将识别精度由 96.25% 和 95% 提高到了 100%; Wang 等[19] 利用 LIBS 结合 PCA 算法和 ANN 算法对不同产地、不同部位的当归、党参、川芎 3 种中药材进行分析鉴定, 达到 99.89% 的识别精度; 郑培超等[20] 利用随机森林分类模型结合 LIBS 技术对石斛进行价格等级分类, 利用袋外数据误差率估计随机森林在不同的决策树个数和分裂属性集中属性个数下的分类效果, 选取最优参数, 将平均识别率提高到了 96.46%.

目前关于 LIBS 结合机器学习算法对人参产地分类还有待研究. 本文基于 LIBS 技术结合机器学习算法对人参产地快速识别, 首先通过 PCA 提取人参样品的 LIBS 光谱数据的特征量, 分别采用 BP 神经网络 (back propagation artificial neural network, BP-ANN) 算法、SVM 算法建立人参产地识别模型, 对东北 5 个产地的同种人参 (白参) 进行聚类分析, 实现了人参产地的识别. 结果表明, LIBS结合机器学习方法是实现人参产地快速识别的有效方法.

2 实验部分

2.1 实验装置

激光诱导击穿光谱技术用于人参产地识别的实验装置如图 1 所示. 激光光源为输出波长 1064 nm, 脉宽 10 ns, 重复频率 10 Hz 的 Nd:YAG 激光器 (Continuum, surellite II), 激光光束直径为 6 mm, 激光光束通过由半波片和格兰棱镜组成的能量调节系统对诱导击穿人参等离子体的脉冲能量进行调控, 激光光束经焦距为 120 mm 的熔石英玻璃平凸透镜聚焦在人参样品表面诱导击穿产生等离子体. 激光光束聚焦焦点位于人参样品表面内 0.8 mm, 目的为避免诱导击穿空气等离子体, 减少对人参光谱分析带来干扰. 在与人参等离子体膨胀轴向方向成 45°的人参等离子体发射光谱方向上, 用焦距为 75 mm 的熔石英透镜收集耦合人参等离子体发射光谱耦合到配有 ICCD 探测器 (1024 × 1024 pixel, DH334) 的中阶梯光栅光谱仪 (Andor, Me5000) 的光纤探头, 光谱仪焦距为 195 mm, 光谱分辨率为 $\lambda/\Delta\lambda \approx 5000$, 一次光谱探测范围为 200—975 nm. 激光器和 ICCD 探测器均由数字脉

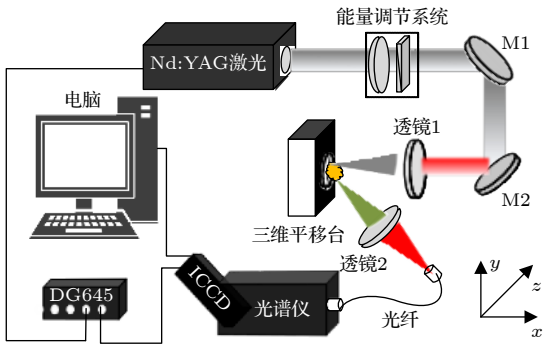

图 1 激光诱导击穿光谱实验装置示意图

Fig. 1. Schematic diagram of the experimental setup of LIBS.

冲延时发生器 (Standoff, DG645) 同步触发工作, 通过优化激光脉冲与 ICCD 探测器间的时间延时和 ICCD 探测器的探测时间门宽, 设定延时和门宽分别为 1 和 5 s, 获得高信背比的人参 LIBS 光谱信号. 为避免人参样品过度烧蚀, 人参样品固定在三维平移台上, 使每个激光脉冲作用在人参样品表面新的位置. 实验中人参 LIBS 光谱为 100 个脉冲进行平均, 降低脉冲能量抖动对人参 LIBS 光谱的稳定性影响. 实验均在标准大气压、室内温度为 22 ℃、空气相对湿度为 25% 的条件下开展.

2.2 样品制备

实验所用的人参样品均为生长年限 15 年的白参, 产地分别为辽宁省石柱 (SZ)、恒仁 (HR), 黑龙江省大兴安岭 (DXAL), 吉林省抚松 (FS)、集安 (JA). LIBS 光谱信号受样品密度、干燥度及研磨均匀性等物理属性的影响, 在实验前先对 5 个产地的人参样品进行纯净、干燥处理, 取干燥后的人参中间支干部位, 使用振动研磨机 (安合盟 (天津) 科技发展有限公司, PrepM-01) 研磨至粉末, 分别经 50 目和 100 目过筛, 取 1.5 mg 样品过筛人参粉末, 使用机械压片机 (安合盟 (天津) 科技发展有限公司, FW-40) 在 25 MPa 压力下压制 25 min, 制成直径 30 mm、厚度为 2 mm 的圆形人参样品, 用于人参产地识别实验样品.

2.3 主成分分析算法

主成分分析 (principal component analaysis, PCA) 算法是一种数据降维的高效信息处理方法, 它采用特征分解获得最大方差的主成分代替原来变量, 可以消除原变量的相关性, 降低数据的维数, 提高建模速度和稳定性. PCA 分析方法为将人参样品 LIBS 光谱的采样值整理并代入向量 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{in})$中 ($n$为光谱特征值), m 为进行降维的 m 组光谱数据, 对样本标准化: 标准化采用 P 维随机变量, 选取 m 个样品, 构造样本阵, 对样本阵进行标准变换:

$$\boldsymbol{Z}_{ij}=\frac{x_{ij}-\overline{x_j}}{\boldsymbol{S}_j}, i=1,2,\cdots,m; j=1,2,\cdots,P, \quad (1)$$

其中 $\overline{x_j}=\dfrac{\sum_{i=1}^{m}x_{ij}}{m}$, $\boldsymbol{S}_j^2=\dfrac{\sum_{i=1}^{m}(x_{ij}-\overline{x_j})^2}{m-1}$, 得到标准化矩阵 $\boldsymbol{Z}$; 通过公式计算相关系数矩阵 $\boldsymbol{R}$:

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1j} \\ r_{21} & r_{22} & \cdots & r_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ r_{i1} & r_{i2} & \cdots & r_{ij} \end{bmatrix}, \quad (2)$$

其中 $r(i,j)=\dfrac{\mathrm{Cov}(i,j)}{\sqrt{\mathrm{Var}[i]\mathrm{Var}[j]}}$, $r(i,j)$ 为 $\boldsymbol{Z}$ 第 i 列和第 j 列的相关系数; 求出协方差矩阵的特征值和特征向量

$$AR=\lambda R, \quad (3)$$

其中, λ 称为 $\boldsymbol{R}$ 的特征值, 非零向量 $\boldsymbol{R}$ 称为 $\boldsymbol{A}$ 对应于特征值 λ 的特征向量; 根据主成分贡献率选择主成分, 计算主成分得分, 将所得主成分作为分类算法的输入参量, 对人参进行产地识别.

2.4 BP 神经网络算法

误差反向传播 (back-propagation algorithm, BP) 神经网络 [21] 是一种按误差逆传播算法训练的多层前馈网络, 它利用大量的数据进行训练获得输入与输出间的映射关系, 再通过梯度下降法不断调整网络的权值和阈值, 使网络的误差达到最小. 图 2 为典型的 BP 人工神经网络结构示意图. 网络 N 个输入节点, L 个输出节点, 隐含层包含 Z 个神经元. $x_1,x_2,\cdots,x_N$ 为网络的实际输入, $y_1,y_2,\cdots,y_L$ 为网络的实际输出.

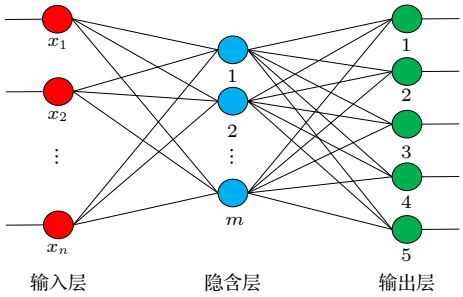

图 2 BP 神经网络结构示意图
Fig. 2. Structure of BP neural network.

BP 神经网络通常由输入层 (input layer)、输出层 (output layer)、一个或多个隐含层 (hidden layer) 组成. 传递函数对误差和训练时间会有很大的影响, 合理地选择传递函数能够降低网络误差, 四种传递函数为 trainlm, trainda, traindm, Traindx. 激活函数以及传递函数的确定需要根据训练数据来进行测试、对比与筛选. 在进行 BP 神经网络仿

真前, 还需要先进行 LIBS 光谱数据的训练集和测试集选择, 从而能够快速实现人参产地鉴定识别.

2.5 SVM 算法

支持向量机[22] (support vector machine, SVM) 实现分类的本质是找一条分割线, 将所有样本点尽可能远离分割线, 即最优超平面. 设训练样本集 $\{(x_i,y_i),i=1,2,\cdots,l\}$, x_i对应样本属性值, y_i对应属性值标签. 对于非线性训练集, 通过一个非线性函数将训练数据x映射到一个高维特征空间, 映射在高维空间中的不同产地人参属性值向量$\phi(x_i)$变为线性可分问题. 此时需构造最优分类超平面并得到决策函数.

分类超平面$f(x)=\omega\cdot\phi(x)+b$, 决策函数$\widetilde{f(x)}=\text{sign}\,[\omega\cdot\phi(x)+b]$. 分类超平面的最优化问题为

$$\min_{\omega,b,\xi_i}\frac{1}{2}\omega^{\mathrm{T}}\omega+C\sum_{i=1}^{l}\xi_i, \tag{4}$$

$$\text{s.t.}\,y_i(\omega^{\mathrm{T}}\cdot x_i+b)\geqslant 1-\xi_i, \tag{5}$$

$$\xi_i\geqslant 0,i=1,\cdots,l,$$

其中C为识别参数, $\xi_i,i=1,\cdots,l$为引入的非负松弛变量. 采用拉格朗日 (Lagrangian) 乘子法求解该问题, 得到对偶形式.

$$\max_{\alpha}\left(\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(X_i,X_j)\right), \tag{6}$$

$$\text{s.t.}\,0\leqslant\alpha\leqslant C,\ i=1,\cdots,l, \tag{7}$$

$$\sum_{i=1}^{l}\alpha_i y_i=0, \tag{8}$$

其中$K(X_i,X_j)=\phi(X_i)^{\mathrm{T}}\phi(X_j)$为核函数, 本实验采用径向基函数 (radial basis function, RBF) 作为核函数, 即

$$K(X_i,X_j)=\phi(X_i)^{\mathrm{T}}\phi(X_j)=\exp(-\|X-X_i\|^2/\sigma^2), \tag{9}$$

式中, σ表示高斯核函数宽度. 最终, 决策函数

$$\widetilde{f(x)}=\text{sign}\left(\sum_{i=1}^{l}y_i\alpha_i K(X_i,X)+b\right). \tag{10}$$

SVM 核心问题是优化惩罚因子C及核函数$g(g=1/\sigma^2)$. 惩罚因子控制对大间隔和最小训练错误率之间的平衡, 用于核空间上非线性可分数据. 本实验基于交叉验证和网格搜索对C与g进行训练, 获得最佳参数C, g进行训练支持向量机算法, 从而能够快速实现人参产地鉴定识别.

3 结果与分析

3.1 特征光谱的选取

进行人参产地识别, 需要考虑实验待测产地人参的 LIBS 全光谱信息, 但 LIBS 全光谱信息量很大, 进而导致机器学习算法计算量过大, 从而人参产地的识别快速性不能得到保证. 为此, 选取合适的特征谱线代表人参样品的全光谱信息, 从而实现快速人参产地识别尤为重要. 激光诱导人参的等离子体发射光谱由线状光谱叠加在连续光谱上组成, 连续背景光谱的存在, 导致了 LIBS 光谱的信背比变低, 本文采用窗口平移平滑法降低背景连续光谱, 5 个产地人参的激光诱导击穿光谱如图 3 所示. 根据美国 NIST 原子光谱数据库对人参 LIBS 光谱进行了元素标记, LIBS 光谱中存在 Mg, Ca, Fe 等矿质营养元素以及 C, H, N, O 等人参组成元素的原子发射光谱. 不同产地人参中元素含量不同, 对应的 LIBS 特征谱线强度有一定的差异, 因而通过多条元素特征光谱强度可对人参产地进行识别. 特征光谱的选择应满足光谱线的重叠少、自吸收现象弱、谱线强度大 (信背比高) 等条件, 最终选取 Mg, Ca, Fe, C, H, N, O 共 7 个元素 8 条特征谱线进行人参产地识别 (特征谱线信息如表 1 所列).

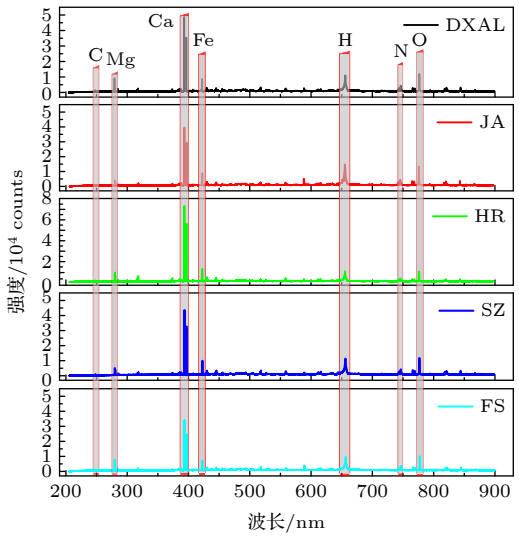

图 3 人参 LIBS 光谱 (产地分别为大兴安岭、集安、恒仁、石柱、抚松)

Fig. 3. LIBS spectra of ginseng(the ginseng origins are DXAL, JA, HR, SZ and FS).

表 1 人参特征谱线及波长

Table 1. Characteristic line and wavelength of ginseng.

元素	波长/nm
C I	247.80
Mg II	279.56
Ca II	393.40; 396.87
Fe I	422.71
H I	656.39
N I	747.07
O I	777.42

在 LIBS 实验过程中, LIBS 光谱强度受到外部气体流动、激光脉冲能量抖动及样品表面元素含量的变化等因素影响, 从而导致在给定实验条件下的 LIBS 光谱强度存在一定的起伏, 这将对依据 LIBS 光谱谱线强度作为元素定量分析产生一定的误差. 因此, 选取 LIBS 光谱中多次重复性实验较为稳定且光谱强度值较大的特征谱线进行 LIBS 光谱强度归一化处理, 能够有效降低外部实验环境等因素造成的 LIBS 光谱强度起伏对定量分析的影响. 本文人参样品 LIBS 光谱中 Ca I 393.40 nm 特征谱线强度最大, 且多次重复实验的光谱强度稳定, 因此选取谱线强度最大的 Ca I 393.40 nm 作为归一化标准. 为降低谱线强度波动对分类结果的影响, 每个 LIBS 光谱中的 8 条特征谱线强度均以 Ca:393.40 nm 光谱强度作归一化处理, 最终得到 5 个产地人参的 657 组数据 (DXAL 117 组、JA 150 组、HR 153 组、SZ 96 组、FS 141 组), 每组数据有 8 个属性, 作为 PCA 的输入: $X_i = (x_{i1}, x_{i2}, \cdots, x_{i8})$.

3.2 主成分分析

由 PCA 分析出人参 LIBS 光谱中 Mg, Ca, Fe, C, H, N, O 共 7 个元素 8 条特征谱线对 LIBS 全谱的主成分贡献情况, 得到前 10 个主成分的贡献率和主成分的累计贡献率如图 4(a) 所示, PC1, PC2 和 PC3 主成分累计贡献率为 92.5%, 可认为 PC1, PC2, PC3 包含了原始人参 LIBS 光谱的大量信息. PC1, PC2 和 PC3 3 个主成分向量组成的三维散点图如图 4(b) 所示. 图 4 中每个散点代表一个人参样本, 可以看出同产地人参样品的特征 LIBS 光谱经 PCA 处理后存在特定的聚集区域, 显示了良好的聚类效果. 结果表明结合 PCA 处理后的 LIBS 光谱数据能够表征人参的产地特征信息, 且能将不同产地人参间的差异进行有效区分. 由图 4(b) 可知, HR, FS 和 DXAL 等产地人参的聚类性较好, 相互之间区分度高, JA 和 SZ 产地人参样品也可聚在一起, 但存在部分重叠.

3.3 结合机器学习对人参产地进行识别

通过 PCA 算法对 5 个人参产地、共 657 组 LIBS 数据进行光谱数据降维处理, 优化 PCA 算法参量, 实现 PC1, PC2 和 PC3 前 3 个主成分累计贡献率为 92.5%, 就以 PC1, PC2 和 PC3 主成分代替人参的 LIBS 特征光谱, 从而构建出人参样品 LIBS 光谱的特征空间向量, 特征向量构成的 657×3 的数据矩阵分别作为 BP 神经网络与 SVM 产地识别算法的输入量, 进而依据 PCA-BP 和 PCA-SVM 算法实现人参产地分类识别. BP 神经网络人参产地识别算法按产地以 2:1 随机选取经主成分降维

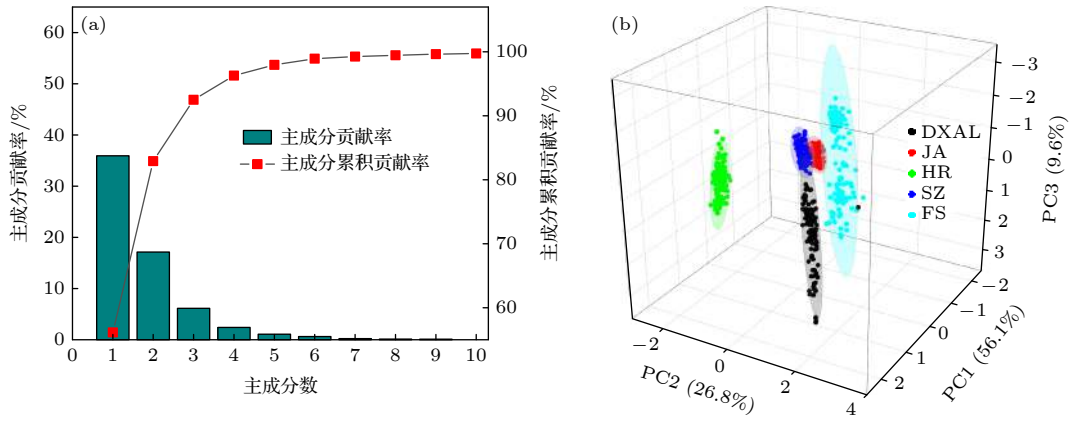

图 4 (a) 各主成分贡献率和主成分累积贡献率; (b) 前 3 个主成分的三维散点图

Fig. 4. (a) Contribution rate of each principal component and cumulative contribution rate of principal component; (b) three-dimensional scatter plot of first three principal components.

处理的 657 组数据, 分为 438 组测试集 (Test) 和 219 组训练集 (Train). 训练集构成的 438 × 3 维数据矩阵作为神经网络训练输入量. 网络的输入向量为三维数据, 因此 BP 神经网络的输入层和输出层的神经元分别为 3 和 5. 运行经多次训练, 最佳隐含层神经元个数为 11, 输入层激励函数为 tansig, 输出层激励函数为 purlin. 网络初始化参数的迭代数设为 1000, 学习率为 0.1, 误差目标为 0.0001.

图 5(a) 为 BP 神经网络最佳验证性能图, 训练误差随训练次数不断减小, 测试均方差 (MSE) 也趋于平缓, 验证曲线 MSE 不再变化时网络训练截止, 网络性能最佳坐标为 (28, 0.03), 达到了最佳网络识别精度. 在此基础上, 以 BP 神经网络机器学习对人参产地分类结果如图 5(b) 所示, 图中“*”表示测试标签, “○”表示实际标签. 当“*”和“○”重合时表明预测准确, 结果显示有 2 个 JA 产地的人参被误判为 SZ 产地, 其他产地 100% 识别, 平均识别精度达到 99.08%, 人参产地识别算法模型运行时间为 2.48 s, 同时结果表明神经网络收敛性良好, 误差个数稳定, 高质量地实现了人参产地判别.

人参产地识别的 SVM 算法的数据选取经主成分降维处理的 657 组数据, 建立与 BP 神经网络算法相同的训练集和测试集, 使用交互检验法优化参数, 得到 PCA-SVM 的网格参数优化如图 6(a) 所示. 图 6(a) 的 x, y 轴分别表示 C, g 取以 2 为底的对数的值, 使用网格搜索方法的分类 (SVC) 参数计算出最佳惩罚因子 C 为 0.14, 最优核函数 g 为 36.76, 此时交叉验证准确率为 99.09%, 训练集准确率为 99.07%. 经参数优化后 SVM 算法对人参产地识别的预测运行结果如图 6(b) 所示. 图 6(b) 中“△”表示预测标签, “○”表示实际标签. 结果表明, 1 个 JA 产地的人参被误判为 SZ, 识别精度为 99.8%. 其他产地的识别精度均为 100%, 平均识别精度为 99.5%, 人参产地识别算法模型运行时间为 14.03 s.

PCA-BP, PCA-SVM 分类算法对人参产地的识别结果如表 2 所列. 由 LIBS 技术结合机器学习的研究结果可知, PCA-BP 和 PCA-SVM 两种分类算法的分类精度均达到了 99% 以上, 实现了目标分类精度, 但在 JA 人参产地的识别上均发生了

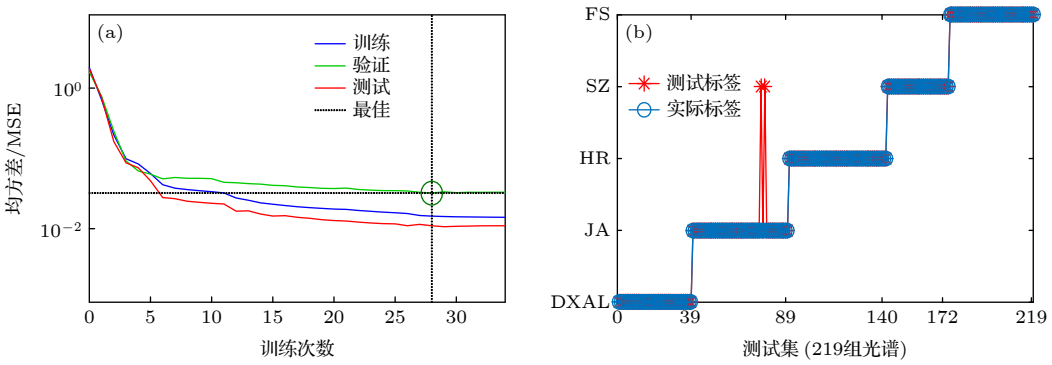

图 5 (a) BP 神经网络训练性能曲线; (b) 分类结果图

Fig. 5. (a) BP neural network training performance curve; (b) classification results.

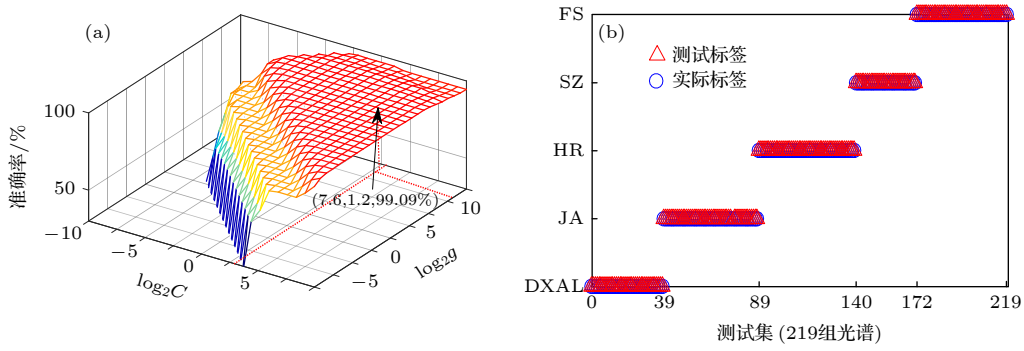

图 6 (a) PCA-SVM 网格参数优化; (b) 分类识别结果图

Fig. 6. (a) PCA-SVM grid parameter optimization; (b) classification recognition result graph.

表 2 人参产地识别结果对比

Table 2. Comparison of ginseng origin identification results.

算法	测试集识别结果		平均识别精度	建模时间/s
	产地	识别精度		
PCA-BP	DXAL	100%	99.08%	2.48
	JA	96%		
	HR	100%		
	SZ	100%		
	FS	100%		
PCA-SVM	DXAL	100%	99.5%	14.03
	JA	98%		
	HR	100%		
	SZ	100%		
	FS	100%		

一定数量的误判. 在算法模型运行时间上, PCA-BP 算法和 PCA-SVM 算法的人参产地识别运算时间分别为 2.48 和 14.03 s, PCA-BP 算法相对于 PCA-SVM 算法的建模速度快了 11.545 s, 有明显优势. 主要原因可能为 BP 神经网络算法具有自主学习能力, 而 SVM 算法需通过核函数将非线性问题实现线性的转化, 识别能力依靠分类超平面的划分, 需寻找最优的核函数以满足识别精度要求, 因而建模时间较 BP 神经网络算法慢.

人参的品质主要由人参皂苷及人参多糖的含量决定, 人参皂苷是固醇类化合物, 人参中皂苷和多糖主要由 C, H, O 等元素决定. 通过分析 5 个产地人参 C I 247.8 nm, H I 656.39 nm, O I 777.42 nm 元素在 Ca II 394.2 nm 元素谱线强度下的归一化强度结果如图 7 所示. 可以看出, JA 和 SZ 两地人参在组成成分上虽因产地的不同导致金属元素的原子发射谱线强度存在差异, 但其 H I 656.39 nm 与 O I 777.42 nm 两条谱线强度的归一化强度几乎相同, 从而导致 JA 和 SZ 人参产地分类时发生误判.

4 结 论

基于激光诱导击穿光谱技术结合机器学习算法对 5 个产地的人参进行了产地的分类识别, 测试集 219 组光谱中, PCA-BP 算法和 PCA-SVM 算法分别正确识别了 217 组和 218 组, 两种算法的识别精度分别为 99.08% 和 99.5%. 但在分类速度上,

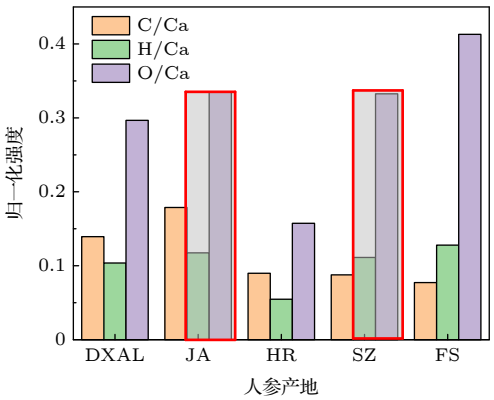

图 7 人参 LIBS 谱中 C, H, O 元素谱线的归一化强度比

Fig. 7. Normalized intensity ratios of C, H and O element lines in the LIBS spectrum.

主成分分析结合神经网络 (PCA-BP) 算法明显优于主成分分析结合支持向量机 (PCA-SVM) 算法. JA 和 SZ 两种人参样本 LIBS 谱线中的 H I 656.39 nm 和 O I 777.42 nm 谱线在以 Ca:393.40 nm 光谱强度作归一化处理后的强度几乎相同, 最终导致两产地发生误判. 实验结果证明, PCA-BP 算法较 PCA-SVM 算法训练速度快, 训练结果较为稳定, 对 5 个产地人参的分类精度较高, 因此利用 LIBS 技术结合机器学习算法可实现人参产地的快速识别.

参考文献

[1] Patel S, Rauf A 2017 *Biomed. Pharmacother.* **85** 120
[2] Kim J H 2018 *J. Ginseng Res.* **42** 264
[3] Huang Y, Gou M J, Jiang K, Wang L J, Yin G, Wang J, Wang P, Tu J S, Wang T J 2019 *Appl. Spectrosc. Rev.* **54** 653
[4] Bec K B, Grabska J, Kirchler C G, Huck C W 2018 *J. Mol. Liq.* **268** 895
[5] Chen J B, Sun S Q, Ma F, Zhou Q 2014 *Spectrochim. Acta, Part A* **128** 629
[6] Fan Q, Chen C, Huang Z, Zhang C M, Liang P J, Zhao S L 2015 *Spectrochim. Acta, Part A* **136** 1621
[7] Radziemski L, Cremers D 2013 *Spectrochim. Acta, Part B* **87** 3
[8] Shao Y, Zhang Y B, Gao X, Du C, Lin J Q 2013 *Spectrosc. Spectr. Anal.* **33** 2593 (in Chinese) [邵妍, 张艳波, 高勋, 杜闯, 林景全 2013 光谱学与光谱分析 **33** 2593]
[9] Costa V C, Augusto A S, Castro J P, Machado R C, Andrade D F, Babos D V, Speranca M A, Gamela R R, Pereira E R 2019 *Quim. Nova* **42** 527
[10] Gottfried J L, Jr F C D L, Munson C A, Miziolek A W 2009 *Anal. Bioanal. Chem.* **395** 283
[11] Tzortzakis S, Anglos D, Gray D 2006 *Opt. Lett.* **31** 1139
[12] Kaiser J, Novotny K, Martin M Z, Hrdlicka A, Malina R, Hartl M, Adam V, Kizek R 2012 *Surf. Sci. Rep.* **67** 233
[13] Gu Y H, Zhao N J, Ma M J, Meng D S, Jia Y, Fang L, Liu J G, Liu W Q 2018 *Spectrosc. Spectr. Anal.* **38** 982 (in Chinese) [谷艳红, 赵南京, 马明俊, 孟德硕, 贾尧, 方丽, 刘建国, 刘文清

2018 光谱学与光谱分析 **38** 982]

[14] Choi J J, Choi S J, Yoh J J 2016 *Appl. Spectrosc.* **70** 1411

[15] Yao M Y, Yang H, Huang L, Chen T B, Rao G F, Liu M H 2017 *Appl. Opt.* **56** 4070

[16] Junjuri R, Gundawar M K 2019 *J. Anal. At. Spectrom.* **34** 1683

[17] Velioglu H M, Sezer B, Bilge G, Baytur S E, Boyaci Il H 2018 *Meat Sci.* **138** 28

[18] Lin J J, Lin X E, Guo L B, Guo Y M, Tang Y, Chu Y W, Tang S S, Che C J 2018 *J. Anal. At. Spectrom.* **33** 1545

[19] Wang J M, Liao X Y, Zheng P C, Xue S W, Peng R 2017 *Anal. Lett.* **51** 575

[20] Zheng P C, Zheng S, Wang J M, Liao X Y, Li X J, Peng R 2020 *Spectrosc. Spectr. Anal.* **40** 941 (in Chinese) [郑培超, 郑爽, 王金梅, 廖香玉, 李晓娟, 彭锐 2020 光谱学与光谱分析 **40** 941]

[21] Koujelev A, Sabsabi M, Ros V M, Laville S, Lui S L 2010 *Planet. Space Sci.* **58** 682

[22] Yu Y, Hao Z Q, Li C M, Guo L B, Li K H, Zeng Q D, Li X Y, Ren Z, Zeng X Y 2013 *Acta Phys. Sin.* **62** 215201 (in Chinese) [于洋, 郝中骐, 李常茂, 郭连波, 李阔湖, 曾庆栋, 李祥友, 任昭, 曾晓雁 2013 物理学报 **62** 215201]

Rapid identification of ginseng origin by laser induced breakdown spectroscopy combined with neural network and support vector machine algorithm*

Dong Peng-Kai [1)] Zhao Shang-Yong [1)] Zheng Ke-Xin [1)] Wang Ji [1)†]
Gao Xun [1)‡] Hao Zuo-Qiang [2)] Lin Jing-Quan [1)]

1) (*School of Science, Changchun University of Science and Technology, Changchun* 130022, *China*)

2) (*School of Physics and Electronics, Shandong Normal University, Jinan* 250358, *China*)

(Received 11 September 2020; revised manuscript received 19 October 2020)

Abstract

Based on laser-induced breakdown spectroscopy and machine learning algorithms, ginseng origin identification model is established by principal component analysis algorithm combined with back-propagation (BP) neural network and support vector machine algorithm to analyze and identify ginseng from five different origins in northeast China (Daxinganling, Ji'an, Hengren, Shizhu, and Fusong). The experiment collects a total of 657 groups of laser-induced breakdown spectral data from five origins of ginseng at 200–975 nm, reduces the background continuous spectrum of the original spectral data by moving window smoothing method, labels the ginseng LIBS spectral elements according to the American NIST atomic spectral database. Eight characteristic spectral lines of 7 elements Mg, Ca, Fe, C, H, N and O are selected for principal component analysis according to characteristic spectral selection conditions. The cumulative contribution rate of the first three principal components of the original spectral data reaches 92.50%, which represents a large amount of information about the original ginseng LIBS spectrum, and the samples show a good aggregation and classification in the principal component space. After dimension reduction, the first three principal components are randomly selected in a ratio of 2 to 1 and divided into 438 test sets and 219 training sets, which are used as the input values of the classification algorithm. The experimental results show that the principal component analysis combined with the BP neural network algorithm and support vector machine algorithm can correctly identify 217 and 218 spectra of 219 spectra of the test set respectively, and the average recognition rate is 99.08% and 99.5% respectively. The modeling time of BP neural network is 11.545 s shorter than that of the support vector machine. Both models misjudged Ji'an Ginseng as Shi zhu ginseng, and the reason for this misjudgment is that the normalized intensity of H and O under Ca element ion emission spectrum are similar due to the proximity of Ji 'an to Shi Zhu in geographical environment. The study presented here demonstrates that laser-induced breakdown spectroscopy combined with machine learning algorithm is a useful technology for rapid identification of ginseng origin and is expected to realize automatic, real-time, rapid and reliable discrimination.

Keywords: laser-induced breakdown spectroscopy, machine learning algorithm, identification of origin, ginseng

PACS: 02.10.Yn, 33.15.Vb, 98.52.Cf, 78.47.dc **DOI:** 10.7498/aps.70.20201520

* Project supported by the National Natural Science Foundation of China (Grant No. 61575030), the Natural Science Foundation of Jilin province, China (Grant Nos. 20180101283JC, 20200301042RQ, 20180201033GX, 20190302125GGX), and the Research Foundation of Education Bureau of Jilin Province, China (Grant No. JJKH20190539KJ).

† Corresponding author. E-mail: jiji_w@163.com

‡ Corresponding author. E-mail: gaoxun@cust.edu.cn