基于 LIBS 与化学计量学的植物叶片分类研究

丁 捷，张大成*，王博文，冯中琦，刘旭阳，朱江峰

西安电子科技大学物理与光电工程学院，陕西 西安 710071

摘 要 激光诱导击穿光谱(LIBS)是一种高效快速的光谱采集手段，可应用于各类物质的元素分析工作中。线性判别分析(LDA)与支持向量机(SVM)是化学计量学中两种常用的有监督算法，均通过对已知不同种类的样本数据进行学习建模，进而实现对未知类别数据的归类。为了实现 LIBS 技术对有机物的高准确率识别，将这两种算法应用到 LIBS 光谱数据的分类中。实验利用波长为 1 064 nm 的纳秒激光烧蚀女贞、珊瑚树、竹子三种植物的叶片，并采集每种树叶 220～432 nm 波段的 100 组光谱数据。通过对 300 组样本的原始光谱数据进行主成分提取，由第一主成分(PC1)和第二主成分(PC2)的得分图得出三种植物光谱的相似度非常高。然后，利用每种叶片 70 组样本的光谱数据作为训练集建模，其余 30 组光谱数据作为测试集来进行树叶种类的预测识别。将 PCA 对原始光谱数据提取得到的前 20 个主成分作为 LDA 与 SVM 建模的属性值。对于 LDA 算法，将属性值分析后得到前两个判别函数值，通过聚类分析发现不同种类的植物叶片光谱数据在空间上的分离效果较好，同一种类基本聚集在一起。再借助马氏距离可得到测试集的平均分类正确率为 96.67%。与此类似，使用 SVM 方法对训练集样本的数据进行学习得到分类超平面，对测试集的平均分类正确率达到 98.9%。研究结果表明，经过 PCA 对数据的预处理，再结合 LDA，SVM 这两种方法可实现 LIBS 技术应用于复杂有机物的快速准确分类，并且 PCA 与 SVM 结合的分类正确率更高。该方法可在食品快速溯源、生物组织原位鉴别、有机爆炸物远程分析等领域应用。

关键词 激光诱导击穿光谱；植物叶片；主成分分析；线性判别分析；支持向量机

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)02-0606-06

引 言

激光诱导击穿光谱(laser-induced breakdown spectroscopy，LIBS)是原子发射光谱技术，将高能量脉冲激光聚焦入射在样品表面上时，可以使样品聚焦区域获得足够高的能量而形成等离子体。随着等离子体的膨胀，等离子体温度快速降低，处于高能级的离子和原子会向低能级或基态跃迁，并辐射出特征谱线。通过特征谱线波长可以确定样品所含元素，同时谱线强度与其所对应的元素含量之间存在定量关系[1]。这就是 LIBS 技术对样品中的元素进行定性和定量分析的基本原理。LIBS 技术是一种消耗样品量极少(亚微克级)、非接触、可实时在线的元素分析手段[2]，目前已经被应用于玉石，液体[3]等物质成分分析领域中。

LIBS 技术直接测量到的主要是元素的特征谱线。对于元素种类组成相似、谱线分布没有明显差异的有机物或复杂样品等难以快速识别。将 LIBS 技术与化学计量学方法结合可以有效识别这些样品的 LIBS 光谱特征。在过去的二十多年里，国际许多研究团队将 LIBS 技术与化学计量学结合开展了大量的物质分类方面的研究工作。法国科学研究中心的 Sirven 等为模拟火星岩石样品的远程识别，将簇类独立软模式(soft independent modeling of class analogy，SIMCA)与偏最小二乘判别分析(partial least squares discrimination analysis，PLS-DA)用于 6 块岩石样品的 LIBS 光谱分类，两种方法的识别正确率分别达到 77.5%和 85.9%。在测试集里加入训练集中不存在的岩石时，二者均表现出了很强的鲁棒性，该结果对于火星的实地探测分析十分重要[4]。西班牙马德里康普顿斯大学的 Moncayo 等测量了多个人的骨骼与牙齿样本的 LIBS 光谱数据，利用骨骼或牙齿样本光谱数据与人工神经网络(artificial neural network，ANN)结合均能实现 95%以上的识别精度。该技术可应用于灾害遇难者的身份识别中[5]。美国麻省理工学院的 Dingari 等对布洛芬，葡萄糖

收稿日期：2020-01-15，**修订日期**：2020-05-08

基金项目：国家自然科学基金项目(11774277)和中央高校基本科研业务费(JB190501)资助

作者简介：丁 捷，1995 年生，西安电子科技大学物理与光电工程学院硕士研究生 e-mail：jdingxd@foxmail.com

*通讯作者 e-mail：dch.zhang@xidian.edu.cn

胺，维生素 C 等药物进行了 LIBS 分类的研究，比较了非线性方法支持向量机(support vector machines，SVM)和其他两种传统的线性方法 SIMCA 和 PLS-DA 的分类结果。研究表明这三种方法对测试集样本的识别准确率都达到 94%以上。但是在鲁棒性测试中，SVM 方法优于 SIMCA 和 PLS-DA 两种方法。此方法可以为假药的鉴别提供新技术[6]。捷克马萨里克大学的 Vítková 等对考古中常见的材料(如贝壳、砖块、陶瓷和骨头等)进行了 LIBS 光谱分析，利用线性判别分析(linear discriminant analysis，LDA)和 ANN 对考古材料进行分类，识别正确率分别为 75%和 87.5%。该方法可以帮助考古研究人员快速分辨现场作业中发现的各种材料碎片[7]。美国特拉华大学的 Celani 等通过手持 LIBS 设备与 K 近邻(k-nearest neighbor，KNN)和 PLS-DA 两种方法结合，实现了对 9 种濒危树种 92%以上的高识别准确率。该技术可以在海关口岸检查濒危树种的非法贸易[8]。近年来，国内研究团队也在 LIBS 分类方面开展了许多研究工作。北京理工大学的王茜倩团队提出了分别利用主成分权重(important weights based on principal component analysis，IW-PCA)和随机森林(random forests，RF)对 LIBS 光谱进行重要性分析，从而提取最优谱线用于分类器输入的方法，然后结合 SVM 对 6 种典型病原菌进行了鉴别，两种模型正确率分别达到 95.79% 和 96.51%[9]。哈尔滨工业大学的李晓晖等采集了 5 种猪肉组织的 LIBS 光谱，利用 KNN 和 SVM 两种方法对脂肪、皮、肌肉达到 99.83%的平均识别率。该结果可为分析临床上微小组织变化、早期病变的诊断提供新方法[10]。有报道利用 SVM 和 PLS-DA 等算法对来自 5 个不同产地的和田玉样品进行了分析，对产地分析的识别正确率分别达到了 99.3%和 97.8%[11]。

以上研究结果表明 LIBS 技术与化学计量学方法相结合是一种在物质分类和产地溯源等领域非常有应用前景的技术。有机物的分类对于食品溯源、爆炸物分析、药品鉴别等诸多领域有着重要的意义。然而在元素组成相似度较高的新鲜有机物识别上，目前研究工作相对有限，分类效果仍有提升空间。本文开展了三种植物树叶(女贞、珊瑚树、竹子)的 LIBS 鉴别工作，探索了将 PCA 分别与 LDA 和 SVM 这两种化学计量学方法结合以提高有机物分类正确率的可行性。

1　实验部分

采用的 LIBS 实验装置如图 1 所示。利用 Nd∶YAG 激光器(Dawa-300，Beamtech，CHN)作为烧蚀光源，激光脉冲宽度为 6 ns，重复频率为 10 Hz，实验中所用脉冲能量为 30 mJ。用石英透镜将激光束聚焦在样品表面，通过一组平凸透镜将等离子体发射光谱收集到光纤中，并传输至双通道光谱仪(AvaSpec-ULS2048-2-USB2，Avantes，NLD)内进行光谱分析。光谱仪的测量范围为 220～432 nm，积分时间为 2 ms。为减少连续辐射谱对元素光谱线的干扰，实验中激光器和光谱仪均由数字信号延迟发生器(DG645，SRS Inc，USA)触发，并将激光脉冲和光谱仪采集之间的延迟时间优化为 300 ns。

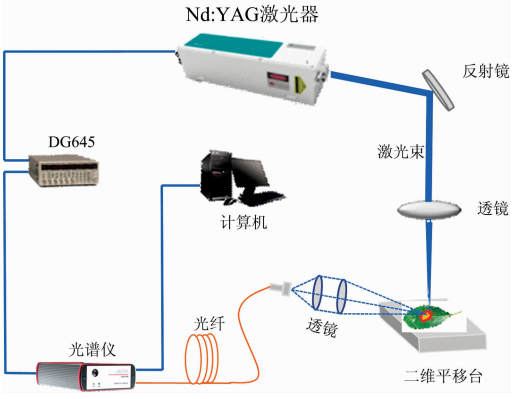

图 1　LIBS 实验装置示意图

Fig. 1　Schematic diagram of LIBS experimental setup

实验选择了西安电子科技大学校园中常见的三种植物(女贞、珊瑚树、竹子)的树叶作为待测样品。用蒸馏水浸泡样品 10 min，以清洗掉树叶表明沉积的灰尘，经自然晾干后粘于样品台上。样品台固定在二维电控位移平台上，按“弓”字形的方式运动。实验测量时，以每片叶子的叶脉为轴，两侧对称采集光谱。每种植物各采集 100 片叶子，1 片树叶只测量 1 组光谱，共得到 300 组光谱数据(女贞、珊瑚树和竹子光谱的 RSD 分别为 23.2%，24.3%和 19.6%)。为降低激光脉冲能量波动对所测得的谱线强度的影响，每组光谱是 100 个激光脉冲作用得到的平均光谱。图 2 为这三种树叶的典型 LIBS 光谱，其谱线形状表现出很高的相似性，难以直接区分。

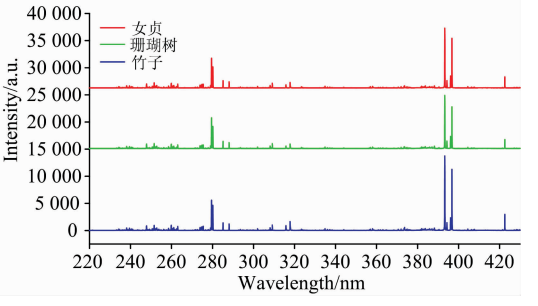

图 2　三种树叶的 LIBS 光谱

Fig. 2　LIBS spectra of three kinds of leaves

2　结果与讨论

2.1　主成分分析

主成分分析(principal component analysis，PCA)是用来研究如何将多指标问题转化为较少综合指标(主成分)问题的方法，这些主成分是传递数据集中包含的主要信息的线性组合，其本质是一种降维的统计过程。PCA 利用正交变换可以将可能具有相关性的谱线数据转化为线性不相关的一组新变量(principal components，PCs)。这种转化令第一主成分具有尽可能大的方差，意味着其包含尽可能多的信息量，并且后续每个成分在保持与前面成分正交的条件下选取方差最大的。

主成分得分图上的散点分布可以表征光谱之间的相似性。图 3 是 300 组光谱数据的第一主成分(PC1)和第二主成分(PC2)的得分图，分别包含了 81.70%和 12.26%的方差信息，代表了原始光谱 93.96%以上的主要信息。可以看出，三种树叶的各自类内聚类效果较为分散，竹子几乎处于另外两种的中间，重叠比较严重。说明三种树叶的光谱数据具有较高的相似性，在元素组成和含量上非常接近。PCA 方法虽然可以很大程度地压缩数据并尽可能保留有效信息，但是难以通过光谱数据的主成分得分图对树叶种类做有效分类。为此，将 PCA 的特征提取作用进一步应用在 LDA 和 SVM 这两种化学计量学方法上，研究这两种方法对三种植物样本种类的识别效果。

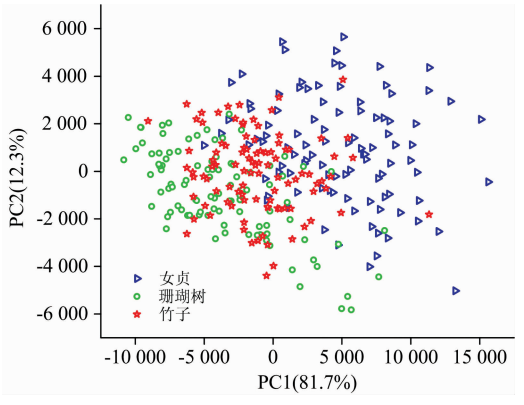

图 3 三种树叶光谱的主成分得分图(前两个主成分)

Fig. 3 Scores of the first two principal components of three kinds of leaves

2.2 线性判别分析与支持向量机

将每种植物叶片的 100 组 LIBS 光谱数据中的 70 组作为训练集，30 组作为测试集，以 PCA 对原始光谱的 2 000 多个谱线数据中提取得到的前 20 个主成分(累计方差大于 99.9%)作为样本属性数据，并为训练集和测试集中同种树叶的每组属性值设定相同的标签。将女贞、珊瑚树、竹子的标签分别标记为 1，2 和 3。由此，训练集中每个样本数据就包括一组属性值和一个标签。分类时先根据训练集中的属性值和标签建立识别模型，然后由模型根据测试集中待测样本的属性值预测出其对应标签，将其与该样本实际对应标签对比得到正确率以检验模型。

2.2.1 线性判别分析

线性判别分析(linear discriminant analysis，LDA)是一种有监督的降维方法，被广泛用于多元统计、模式识别和机器学习等领域中。其基本思想是将高维的样本数据投影到最佳鉴别的低维矢量空间，以达到抽取分类信息和压缩特征空间维数的目的。该算法的核心则在于寻找最能区分不同类数据的最佳投影方向，使得类间距离与类内距离的比值最大化。数据经过处理后在空间上表现出同一种类别数据的投影点尽可能接近，而不同类别的数据的投影点相互远离的趋势。

随机选取每种树叶的 70 组光谱数据用于建立判别模型。训练集中的每组光谱数据经过模型分析后可得到一系列判别函数值，利用前两个判别函数值可作出如图 4 所示的散点图。与图 3 相比，图 4 中三种植物树叶的聚类效果更好，不同类样本数据之间的间隔也更为明显，未出现某个样本处于其他种类植物树叶样本聚集区域的情况。由于 LDA 可以使所获得的新数据中同种树叶的数据相似性提高，不同种植物树叶数据差异扩大。相对于 PCA 仅仅处理了数据间的相关性和冗杂性，LDA 的判别能力更强。通过得到的判别模型对剩余每种树叶的 30 个样本(共 90 组数据)进行验证分析，利用前两个判别函数值可作出图 5。可以看出同一种类的样本也都各自聚在一起，仅有个别不同种类的样本间距较近。其中 1 个竹子样本非常靠近女贞样本群的边缘，2 个竹子样本几乎处于三种样本的交界中心。此外，还有 2 个珊瑚树样本离珊瑚树样本群和竹子样本群的远近程度相当。

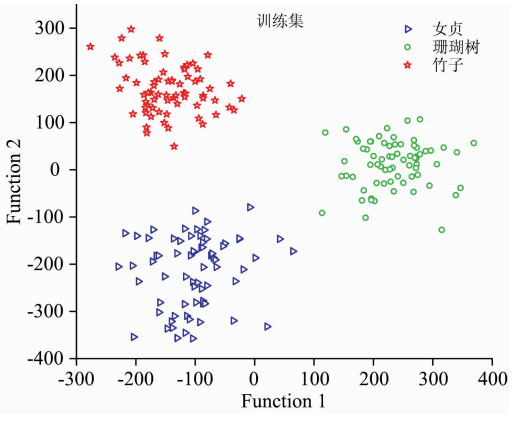

图 4 训练集样本的前两个判别函数的散点图

Fig. 4 The scatter diagram of the firsttwo discriminant functions of training set samples

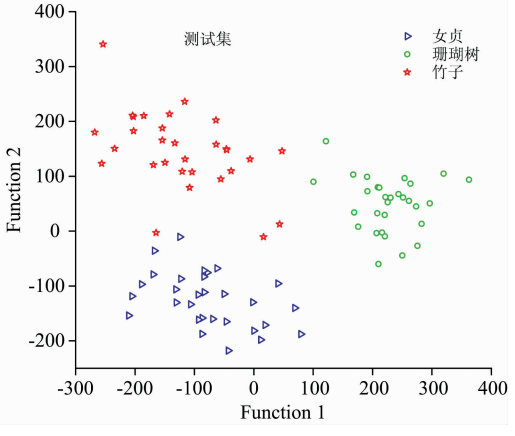

图 5 测试集样本的前两个判别函数的散点图

Fig. 5 The scatter diagram of the first two discriminant functions of test set samples

为了定量表征测试集中的待测样本属于各种类的可能性，引入马氏距离定量描述测试集中的未知样本与训练集每种样本群的“相似程度”，将未知样本划入与其相似性最高的类别。马氏距离是一种无量纲、与变量尺度无关且考虑了数据集相关性的广义距离，它可以用来测量任一样品点 A 与某

一样品集 P 之间的距离。其在计算过程中引入协方差矩阵，使得实验中均值较高的变量影响减小，同时均值较低的变量影响增大，最终令所有变量对分类的贡献趋于一致。为了明确各个样品所属分类，分别计算测试集中所有待测试样本与三种植物叶片训练集样本群的马氏距离，马氏距离越小，说明其与对应类别相似性越高，反之相似性越低。如图 6 所示，蓝色，绿色和红色标志分别表示该待测样本与训练集中女贞、珊瑚树和竹子样本集的马氏距离。从图 6 中可以看出待测样本与其实际类别对应样本集的马氏距离大多接近于 0，并且另外两个马氏距离远大于 0，距离差异很大，即相似性差异明显。这表明经过 LDA 处理，光谱数据按类别在空间上完全分离开来。因此，选择 3 个马氏距离中最小值所对应的类别作为该样品的预测类别。根据马氏距离计算得到测试集分类结果如图 7 所示，测试集中 30 个女贞样本(1—30)和 30 个珊瑚树样本(30—60)全部分类正确；竹子样本(60—90)中有 2 个被误分为女贞，1 个被误分为珊瑚树。最终，在 90 个测试集样本中正确分类 87 个，平均正确率达到 96.67%。

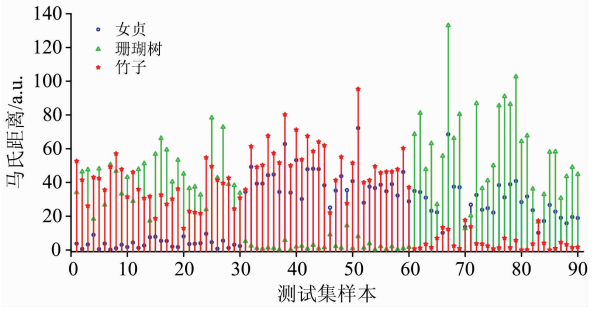

图 6　每个测试集样本的 3 个马氏距离

Fig. 6　Three Mahalanobis distances of each test set sample

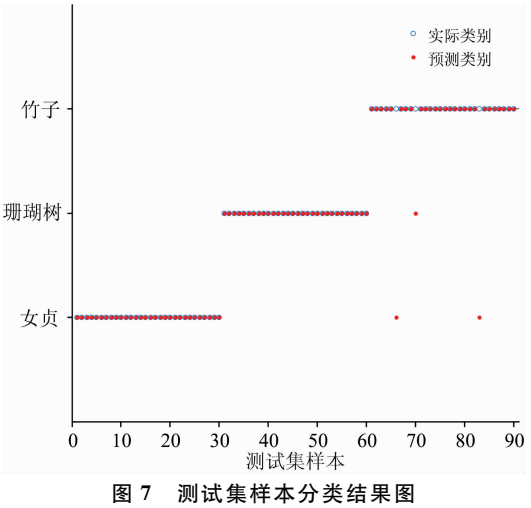

图 7　测试集样本分类结果图

Fig. 7　Classification results of test set samples

2.2.2　支持向量机

支持向量机(support vector machines，SVM)是 Cortes 和 Vapnik 提出的一种较新的非线性分类方法[14]。SVM 是定义在特征空间上的间隔最大分类器，通过将数据映射到在高维空间，利用两类间距离最近的训练点(支持向量)求得一系列对两类分割的超平面。而所求的最佳超平面距两类的支持向量一样远，使得不同类数据之间的分布间隔最大化，其本质是一种二分类模型。在应对多类问题时，采取“一类对其余”的方法，每次仍然解一个二分类的问题。SVM 在分类数据集时不存在必须线性可分的限制，在解决小样本、非线性及高维模式识别中表现出许多特有的优势。

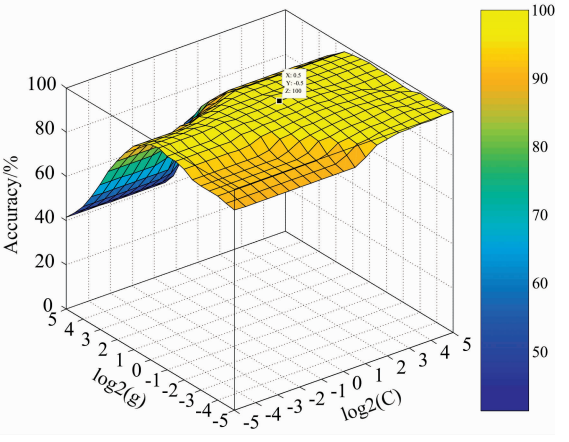

图 8　SVM 参数寻优图

Fig. 8　SVM parameter optimization

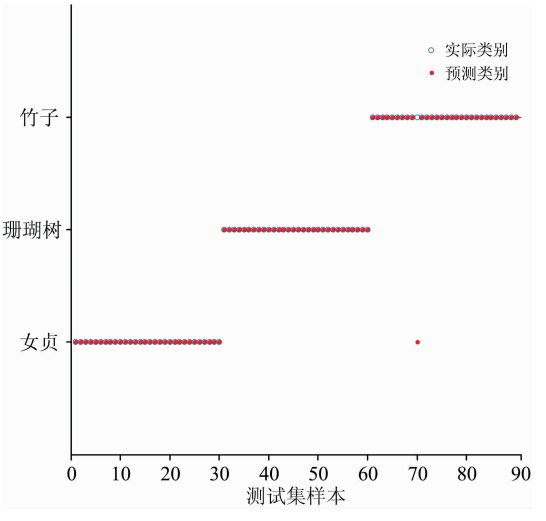

图 9　SVM 分类结果图

Fig. 9　Classification results of SVM

本工作使用了基于 MATLAB 的 Library for Support Vector Machines (LISVM)工具箱[13]对三种植物叶片的 LIBS 光谱建模。其中，核函数为径向基函数(RBF)，由于工具箱中惩戒因子 C，核参数 g 会直接影响对最优分类超平面的求解，因此，需要对 C，g 进行参数寻优才能建立更准确的 SVM 分类器模型。采用交互验证法寻找最佳(C，g)。考虑到建模时间和准确性，将 C 与 g 的调节范围均设置为(2^{-5}，2^{5})，参数的步进值设为 $2^{0.5}$。图 8 是(C，g)的参数寻优图，从图 8 可以看出不同(C，g)对应的训练集平均正确率，程序给出的最优参数(C，g)为($2^{0.5}$，$2^{-0.5}$)，对应最高训练集正确率为 100%。利用该参数对应的分类器可对测试集每组属

性值的类别进行预测。图 9 给出了测试集中 90 个待测样本的预测类别与实际类别，竹子叶中有 1 个被误分为女贞树叶，而这个样本在 LDA 中被误分为珊瑚树。结果中共 89 个正确分类，测试集的平均正确率达到 98.89%。

3 结 论

采集了女贞、珊瑚树和竹子三种植物叶片在 220～432 nm 波段的 LIBS 光谱。利用 PCA 对三种植物叶片的光谱数据进行可视化分析，在得分图上得到的样本点重叠严重，难以实现女贞、珊瑚树、竹子的准确识别。将 PCA 提取的前 20 个主成分输入 LDA 和 SVM 模型进行三种植物叶片光谱数据的分类。在测量结果中，LDA 结合马氏距离时，测试集 90 个待测样本对训练集中各类样本集的距离差异明显，仅对 3 个竹子样品分类错误，女贞与珊瑚树样品全部正确归类，平均正确率达到 96.67%；SVM 经过参数寻优后，得到的最优参数对应的模型在训练集中得到了 100%的分类正确率，而对于测试集样本，仅有 1 个竹子叶片被误分，平均正确率为 98.89%。研究表明，将 PCA 与 LDA、SVM 这两种有监督的化学计量学方法结合能够实现对新鲜植物样品 LIBS 光谱的准确识别，并且 PCA 与 SVM 结合的分类结果优于 PCA 与 LDA 方法结合的分类结果。该方法有助于 LIBS 技术在食品快速溯源、生物组织原位鉴别、有机爆炸物远程分析等领域应用。

References

[1] Senesi G S, Senesi N. Analytica Chimica Acta, 2016, 938: 7.
[2] Wang Z, Yuan T B, Hou Z Y, et al. Frontiers of Physics, 2014, 9(4): 419.
[3] Zhang D C, Hai B, Zhu J F, et al. Optics Express, 2018, 26(14): 18794.
[4] Sirven J B, Salle B, Mauchien P, et al. Journal of Analytical Atomic Spectrometry, 2007, 22(12): 1471.
[5] Moncayo S, Manzoor S, Ugidos T, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2014, 101: 21.
[6] Dingrai N C, Barman I, Myakalwar A K, et al. Analytical Chemistry, 2012, 84(6): 2686.
[7] Vítková G, Novotný K, Prokeš L, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2012, 73(3): 1.
[8] Celani C P, Lancaster C A, Jordan J A, et al. Analyst, 2019, 144(17): 5117.
[9] Wang Q Q, Teng G, Qiao X L, et al. Biomedical Optics Express, 2018, 9(11): 5837.
[10] Li X H, Yang S B, Fan R W, et al. Optics and Laser Technology, 2018, 102: 233.
[11] Yu J L, Hou Z Y, Sheta S, et al. Analytical Methods, 2018, 10(3): 281.
[12] Cortes C, Vapnik V. Machine Learning, 1995, 20(3): 273.
[13] Chang C C, Lin C J. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1.

The Classification of Plant Leaves by Applying Chemometrics Methods on Laser-Induced Breakdown Spectroscopy

DING Jie, ZHANG Da-cheng*, WANG Bo-wen, FENG Zhong-qi, LIU Xu-yang, ZHU Jiang-feng
School of Physics and Optoelectronic Engineering, Xidian University, Xi'an 710071, China

Abstract Laser induced breakdown spectroscopy (LIBS) is a highly efficient and rapid elemental analysis method. It can be applied to the elemental analysis of various materials. Linear discriminant analysis (LDA) and support vector machine (SVM) are two commonly used supervised algorithms in chemometrics. These two methods both need to build the models with known sample data, and then to classify unknown sample data. In order to achieve high accuracy of recognition for organics by LIBS technology, these two algorithms were used to analyze LIBS spectra. In this experiment, a nanosecond laser with 1 064 nm wavelength was used to ablate three kinds of plant leaves (Ligustrum lucidum, Viburnum odoratissinum, bamboo) to produce plasma. The plasma spectra were acquired by an optical fiber spectrometer in the range of 220 to 432 nm. 100 spectra from each kind of plant leaves were collected. Firstly, the principal component extraction for the original spectral data of 300 samples was carried out. Then the first two principal components (PC1, PC2) were used to make the score plot. The spectra of these three kinds of plant leaves are very similarities so that they could not be identified directly. Then, 70 spectra of each kind of plant sample were set as a train set, and the other 30 spectra were used as the test set to test the classification model. The first 20 principal components extracted by the PCA were used as attribute values for modeling of LDA and SVM. For the LDA, the spectra were processed to obtain the first two discriminant function values. The larger scatter distribution intervals for different types of leaves can be acquired by plotting the discriminant function values. Then combined with the Mahalanobis distance, the

average classification accuracy of the test set was up to 96.67%. Similarly, the SVM method was used to learn the characters of the train set to obtain the classification hyperplane. The average classification accuracy rate of SVM for the test set was up to 98.89%, which is better than LDA. This work can be helpful to food traceability, in situ identification of biological tissues and remote analysis of organic explosives by LIBS technology.

Keywords Laser induced breakdown spectroscopy; Plant leaves; Principal component analysis; Linear discriminant analysis; Support vector machine

(Received Jan. 15, 2020; accepted May 8, 2020)

* Corresponding author