基于波段选择的拉曼光谱血痕鉴别

杨志超[1,2]，石 璐[1]，蔡 竞[1]，张 辉[1]

1. 浙江警察学院，浙江 杭州 310053
2. 毒品防控技术浙江省重点实验室，浙江 杭州 310053

摘 要 血痕的种属鉴别在刑事技术和检验检疫等领域有重要的实践意义，拉曼光谱技术为血痕种属鉴别提供了思路。实验采集人血及猪、鸡、鸭、牛、鼠5种动物的血样并获取其拉曼光谱，采用Savitzky-Golay方法平滑降噪，airPLS方法进行基线校正，选取100～1 700 cm^{-1}波段进行实验。训练集有600组数据，测试集有300组拉曼光谱数据。第一部分实验对比了PLS-DA，LDA，PCA+LDA，SVM和PCA+SVM等方法，测试集准确率分别为84.0%，49.3%，78%，83.0%和85.7%，验证了降维算法结合SVM分类器的有效性。第二部分采用互信息算法、遗传算法和等间隔组合三种波段选择算法，结合SVM分类器做对比实验，结果显示互信息结合SVM算法的分类准确率最优，在选择波段数为50时，测试集准确率达到86.0%。在波段选择数为300时，三种波段选择算法结合SVM分类器的准确率都达到93%左右，大幅高于传统分类方法。实验结果表明，采用波段选择算法进行光谱降维，可以有效的提高算法的准确率和鲁棒性，同时使拉曼光谱种属鉴定的可解释性更强。波段选择算法确定了血痕鉴别的关键波段位置，对设计用于执法的便携式拉曼系统也有重要意义。

关键词 血痕；拉曼光谱；分类模型；波段选择

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)10-3137-05

引 言

在公安刑侦、检验检疫等工作中，血痕是重要的物证之一，往往需要对血痕进行种属鉴别，判断其为何物种所留。传统的血痕鉴别技术有酶免疫分析、DNA检测法、高效液相色谱法等[1-3]，此类方法会损耗检测样品，因物证样本是行政执法和法庭审判的重要证据，应尽量采用无损的检测方法。拉曼光谱技术具有无损、无需前处理、分析检测速度快的优势，特别适合物证检验。特别是表面增强拉曼技术[4-7]，分子附着与纳米金属材料表面，通过纳米金属颗粒放大待测物的拉曼光谱信号，增强倍率可达10^6以上。目前，拉曼光谱技术及其相关技术已被广泛应用于化工、医学、半导体、地质等领域[8-10]，在血痕种属鉴别等相关领域也日渐成熟。

利用拉曼光谱技术可以快速有效的鉴别血痕的物种归属。白鹏利等人以3种不同物种动物和人类血痕为研究样本，采取拉曼光谱技术结合主成分分析法(PCA)，对于样本血痕进行定性识别[11]。郑祥权等采用了人血与比格犬血作为实验样本，检测血痕样本的拉曼光谱数据，结合PCA和线性判别分析(LDA)分类算法，构建了一种可以对人血和犬血进行种属判断的多元统计算法模型，测试集的准确率达到90%[12]。董家林等利用海洋光学Raman光谱仪测得共326例样本数据(人110例、犬116例、兔100例)，采用SVM分类器，训练集分类正确率达100%，测试集分类正确率达93.52%[13]。

本文收集人血和鸡、鸭、猪、牛、鼠5种动物血痕的拉曼光谱数据，对数据完成降噪和基线矫正，结合特征选择算法，建立多分类模型，对比各特征选择算法和分类模型对准确率的影响，并对结果进行验证，建立一套血痕种属鉴别的快速检测方法。

1 实验部分

1.1 血痕样本

收集鸡、鸭、猪、牛、鼠5种动物的血痕样本75份，每种动物的血痕样本15份。收集15名健康志愿者血痕样本15份，所有志愿者同意协助完成实验。所有血痕样本不做任何的前处理，取血后24 h内测量，血液滴载玻片表面，静置约

收稿日期：2020-08-20，**修订日期**：2020-12-20

基金项目：国家重点研发计划项目(2018YFC0807401)，浙江省教育厅科研项目(Y201737880)，浙江警察学院项目(2020XJY015)资助

作者简介：杨志超，1985年生，浙江警察学院讲师 e-mail：yangzhichao@zjjcxy.cn

2 h，待血液完全凝固后获取其拉曼光谱，实验环境温度为 20 ℃，湿度为 40%。

1.2 训练集、验证集和测试集的划分

将 6 类物种，每类物种 15 个样本分成训练集和测试集。每类物种随机选取 10 个样本作为训练集，利用训练集样本完成模型的建立和调参。另外 5 个样本作为测试集，利用测试集数据做最终的模型评价。对每个样本随机选取 10 个不同的位置获得拉曼光谱数据。最终得到的训练集中有 600 组拉曼光谱数据，测试集中有 300 组拉曼光谱数据。

1.3 拉曼光谱仪与计算环境

实验采用美国 Thermo Fisher 公司生产的 DXR2xi 显微激光拉曼成像光谱仪，拉曼光谱仪具有超低暗噪声，单光子信号探测器等优势。计算机环境为 Intel(R) Core(TM) i5-5200U CPU @ 2.2 GHz，RAM：12.0 GB，64 位操作系统。

1.4 拉曼光谱数据的获取与校正

利用拉曼光谱仪获取血痕的拉曼光谱，实验考察了不同的激发波长、物镜倍数、激光强度、曝光时间、扫描次数等采集参数，综合比较对样本的破坏、荧光干扰、拉曼信号信噪比、实验效率等方面。实验选择 633 nm 激光作为激发光源，采用 10×物镜聚焦，激光强度为 3.0 mW，曝光时间为 0.2 s，扫描次数为 100 次，采集后的拉曼光谱的信噪比约 40。实验采用迭代自适应加权惩罚最小二乘法校正基线，使用 S-G 平滑滤波实现平滑处理，选取 100～1 700 cm^{-1} 波段测试研究，共 830 个波段。

2 结果与讨论

研究分两部分实验，第一部分实验，建立 PLS-DA，LDA，PCA+LDA 与 SVM，PCA+SVM 模型进行对比实验，比较 SVM 分类方法相对于其他两种方法的准确率，以及 PCA 降维的效果。第二部分实验，采用三种波段选择方法对拉曼光谱降维，将被选择的波段数据放入 SVM 分类器中，探讨波段选择方法对分类准确率的影响。

2.1 分类方法实验

线性判别分析(linear discriminant analysis，LDA)是一种多元线性学习方法，思路是将数据投影到一条直线上，使不同类数据的投影之间的距离尽量远，且同类数据的投影之间的距离尽量近。偏最小二乘判别分析(partial least squares discriminant analysis，PLS-DA)是一种用于多元判别分析方法，适用于样本少、特征多，且特征变量之间存在多重共线性的情况。实验通过十折交叉验证，对参与建模的前 10 个主成分做判别分析并计算准确率。

支持向量机(support vector machine，SVM)利用核函数把样本从低维空间映射到高维空间，寻找最优超平面将特征空间划分开。只有少量的支持向量在 SVM 分类中起决定作用，不仅避免了“维数灾难”问题，也使 SVM 算法鲁棒性更强。因此，SVM 算法适用于小样本、高维度的拉曼光谱数据问题。SVM 分类模型有两个重要的参数 C 和 gamma。通过网格搜索的方法确定最佳的 C 和 gamma 组合，如图 1 所示，当 $C=100$，gamma=0.001，准确率达到 90%以上。

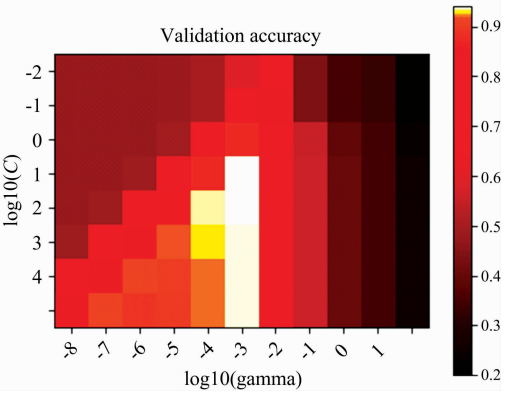

图 1 C 和 gamma 网格搜索的结果

Fig. 1 Results of grid search for C and gamma

表 1 五种分类方法在训练集和测试集的准确率

Tabel 1 Accuracy of training set and test set for five methods

分类方法	训练集准确率/%	测试集准确率/%
PLS-DA	87.2	84.0
LDA	54.5	49.3
PCA+LDA	84.5	78.0
SVM	92.3	83.0
PCA+SVM	89.5	85.7

不同的分类方法的准确率结果如表 1 所示，PCA+SVM 方法在测试集中的准确率最高，达 85.7%。LDA 的准确率最低，可能的原因是血痕的拉曼光谱数据存在严重的共线性问题。通过 PCA 降维后，LDA 和 SVM 算法在测试集中的准确率都有所提高，可以说 PCA 降维对于测试准确率的提高有一定的帮助。另一方面，SVM 算法的准确率相对于 LDA 和 PLS-DA 都更高。因此，SVM 分类器是更优的选择。

2.2 波段选择实验

光谱数据的降维方法分为基于数学变换的降维方法和基于波段选择的降维方法。基于数学变换的降维方法，比如 PCA，LDA 和 PLS 等，改变了原始数据的物理意义，可解释性差，同时复杂的降维算法也增加了计算成本。基于波段选择的降维是从原始光谱数据中筛选出波段子集，剔除不起作用或有干扰作用的冗余波段，不会改变原始特征数据，不产生新的特征，所挑选出来的波段依然保持原来的物理意义，可解释性强，并且有效的提高计算效能。如图 2 所示，t1，t2，t3，…表示原始的拉曼光谱数据，s1，s2 表示经过数学变换降维后的光谱数据。

互信息(mutual information，MI)度量了两个随机变量之间的统计依赖关系，因此可以用来评估每个波段对分类的相对效用。相对于单独使用信息熵来说，互信息搭起了波段信息与实际目标之间的关系。计算每一个波段与类别信息之间的互信息值，然后对波段的互信息进行降序排列，选择出互信息值最大的前 K 个波段，组成波段子集。

遗传算法(genetic algorithm，GA)是一种模拟生物遗传机理的模型，通过适者生存的方式寻找最优解。从一个随机的种群开始，逐代演化出更近似的解。依据对问题的适应性

来选择个体，然后个体之间进行交叉和变异产生新的种群。实验中遗传算法主要参数：变异概率 2%，迭代次数 150 次，种群个体数为 200。

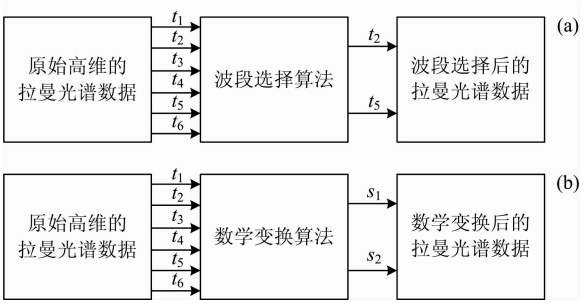

图 2　光谱降维算法示意图

(a)：波段选择；(b)：数学变换

Fig. 2　Schematic diagram of a spectral dimensionality reduction algorithm

(a)：Band selection method；(b)：Mathematical transformation method

等间隔组合法(equidistant combination，EC)本质上是降低了光谱的分辨率，达到波段选择的目的。其主要思想是在一定光谱范围内以相同的间隔提取波段。等间隔组合方法的参数包括以下三个：起始波长、波长个数、相邻波长点之间的间隔数，比如(101，200，5)的波段数为 20。本实验中，波段数相同的，取准确率最高者。

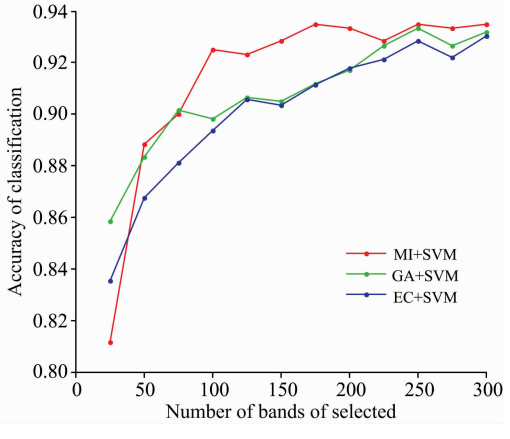

图 3　不同的波段数量下三种波段选择方法的训练集准确率

Fig. 3　The training set accuracy of the three band selection methods for different bands number

训练数据集中，采用 SVM 分类器，三种波段选择方法 10 折交叉验证的准确率表现如图 3 所示。在取 25 个波段时，互信息法准确率不高，随着波段的数量增加，互信息法所选择的波段准确率提升较快，并且在 150 个波段后保持稳定，整体准确率较高。在波段选择数为 300 时，遗传算法、等间隔组合算法与互信息法的准确率接近，达到 93%左右。

在选择波段的数量为 50 时，MI＋SVM，GA＋SVM，EC＋SVM 在训练集准确率为 88.8%，88.3%和 86.8%，已达到 PCA＋SVM 方法的相近的水平。根据训练集中确定的最优光谱波段，同样的选取测试集中对应的 50 个波段组合，并放入 SVM 分类器中验证方法的可靠性，结合 PCA＋SVM 和 PLS-DA，结果对比如图 4 所示。

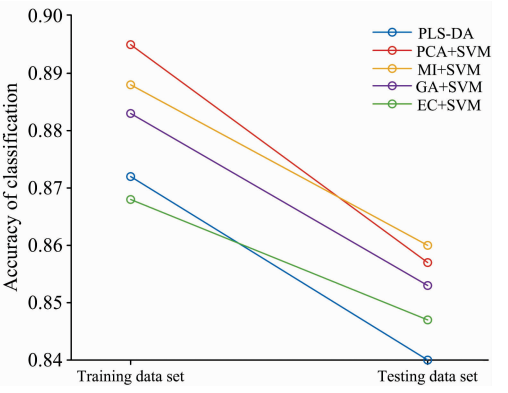

图 4　五种方法在训练集和测试集的准确率对比

Fig. 4　Accuracy of training set and testing set for five methods

在选择波段的数量为 50 时，MI＋SVM，GA＋SVM，EC＋SVM 在测试集准确率为 86.0%，85.3%和 84.7%，互信息法过滤得到的 50 个波段组合，在训练集和测试集准确率都是最高的。在测试集数据中，MI＋SVM 算法的分类准确率高于 PCA＋SVM(86.7%)。另外，从图 4 中可以发现，采用波段选择降维方法，训练集准确率与测试集准确率之差更小，主要原因是波段选择方法排除了冗余的干扰波段的影响，其表现更加稳定。

未参与建模的 300 组测试样本(6 个物种各 50 组拉曼光谱)中，人、猪、牛、鼠、鸡和鸭的血液拉曼光谱预测准确率分别为 84%，80%，84%，82%，92%和 94%。人血与猪血之间是判错率较高，人血的错例中有 87.5%是被错判为猪血的，猪血的错例中有 60%是被错判为人血的，这与猪血与人血的拉曼光谱更为相近有关。鸡血和鸭血之间容易混淆，两者同为禽类，拉曼光谱更为相似。禽类(鸡、鸭)和哺乳类动物(人、猪、牛、鼠)之间是完全没有错判的，说明禽类和哺乳类动物血液的拉曼光谱有较大区别。

根据图 3，在最优波段数量为 25 时，分类准确率达到了 80%以上。图5表示了互信息法所得到的最优的25个拉曼

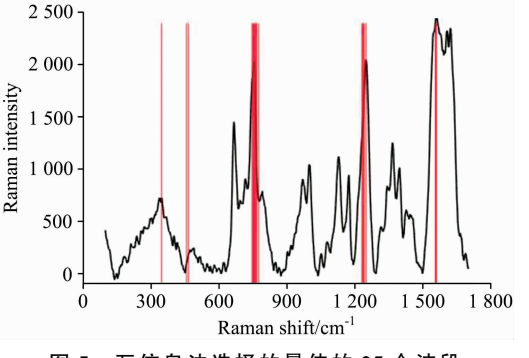

图 5　互信息法选择的最佳的 25 个波段

Fig. 5　The best 25 bands obtained by mutual information method

波段组合，红色的线段代表了所选择的波段。25 个波段主要集中在 755，1 230 和 1 560 cm^{-1}附近。其中可知的是 1 230 cm^{-1}是 C═S 键引起的，1 560 cm^{-1}是 C═C 或 N═N 键引起的，主要是苯丙氨酸、络氨酸及色氨酸等氨基酸所对应的拉曼光谱[15]，说明不同物种血液中氨基酸的多样性可以通过其拉曼光谱反映出来。物种血液中核酸碱基含量的差异，也会使拉曼谱峰相对强度改变。

3 结 论

在血痕种属鉴别方面，以 SVM 算法作为拉曼光谱数据的分类器，相对于 LDA 和 PLS-DA 分类器的准确率更高。波段选择降维方法应用于血痕拉曼光谱鉴别充分体现出其有效性。通过互信息法过滤得到的最佳波段组合，再利用 SVM 算法分类，其在验证集和测试集准确率都是较高的。在选择 50 个波段时，分别达到 88.8%和 86.0%。PCA＋SVM 算法的准确率低于 MI＋SVM。波段选择方法的适应性更好、可解释性更强，对利用拉曼光谱在其他领域应用有借鉴意义。在实践方面，波段选择可以简化拉曼光谱系统，使该技术应用于刑事技术、海关检疫等方面更加快捷和经济。

References

[1] Toishi Y, Tsunoda N, Nagata S, et al. Journal of Reproduction & Development, 2018, 64: 41.
[2] Li Y, Pan X, Roberts M L, et al. Epigenomics, 2018, 10: 797.
[3] Kotnik P, Krajnc M K, Pahor A, et al. Journal of Pharmaceutical & Biomedical Analysis, 2018, 150: 137.
[4] ZHU Zhi-hui, MENG Fan-hao, XIA Jia-bin, et al(朱智慧，孟凡皓，夏嘉斌，等). Acta Academiae Medicinae Sinicae(中国医学科学院学报), 2020, 42(3): 399.
[5] Ding Qianqian, Wang Jing, Chen Xueyan, et al. Nano Letters, 2020, 20(10): 7304.
[6] Ceballos M, Arizmendi-Morquecho A, Sanchez-Dominguez M, et al. Materials Chemistry and Physics, 2020, 240: 122225.
[7] Sarychev A K, Bykov I V, Boginskaya I A, et al. Optical and Quantum Electronics, 2020, 52(1): 26.
[8] LIN Ling, ZHU Zhan-yuan, WANG Lan(林 玲，朱占元，王 兰). Guangdong Chemical Industry(广东化工), 2019, (2): 92.
[9] Wang Z, Yang H, Wang M, et al. Colloids & Surfaces A Physicochemical & Engineering Aspects, 2018, 546: 48.
[10] Lee M, Kim H, Kim E, et al. ACS Applied Materials & Interfaces, 2018, 10: 37829.
[11] BAI Peng-li, WANG Jun, YIN Huan-cai, et al(白鹏利，王 钧，尹焕才，等). The Journal of Light Scattering(光散射学报), 2016, 28(2): 163.
[12] ZHENG Xiang-quan, LIAO Xin, XU Yi, et al(郑祥权，廖 鑫，徐 溢，等). Chemical Journal in Chinese Universities(高等学校化学学报), 2017, 38(4): 575.
[13] DONG Jia-lin, HONG Ming-jian, ZHENG Xiang-quan, et al(董家林，洪明坚，郑祥权，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(2): 459.
[14] DU Pei-jun, XIA Jun-shi, XUE Chao-hui, et al(杜培军，夏俊士，薛朝辉，等). Journal of Remote Sensing(遥感学报), 2016, 20(2): 236.
[15] Feng Pengmian, Feng Lijing. International Journal of Biological Macromolecules, 2020, 162: 931.

Raman Spectral Blood Stain Identification Based on Band Selection

YANG Zhi-chao[1,2], SHI Lu[1], CAI Jing[1], ZHANG Hui[1]

1. Zhejiang Police College, Hangzhou 310053, China
2. Key Laboratory of Drug Prevention and Control Technology of Zhejiang Province, Hangzhou 310053, China

Abstract The species identification of blood stains has important practical significance in criminal technology and inspection and quarantine. Raman spectroscopy provides an idea for the identification of bloodstain species. In this paper, human blood samples and blood samples of pig, chicken, duck, cow and mouse were collected and their Raman spectra were obtained. Savitzky-Golay method was used to smooth noise reduction, airPLS method was used for baseline correction, and 100～1 700 cm^{-1} bands were selected for the experiment. The training set contained 600 sets of data, and the test set contained 300 sets of Raman spectral data. The first part of the experiment compared plS-DA, LDA, PCA＋LDA, SVM and PCA＋SVM. The accuracy of the test set was 84.0%, 49.3%, 78%, 83.0% and 85.7% respectively, which verified the effectiveness of the combination of the dimension-reduction algorithm and the SVM classifier. In the second part, three band selection algorithms of mutual information algorithm, genetic algorithm and equispaced combination were adopted. A comparative experiment was conducted in combination

with the SVM classifier. The results showed that the combination of mutual information and the SVM algorithm had the best classification accuracy. When the number of band selection is 300, the accuracy of the three band selection algorithms combined with the SVM classifier is about 93%, which is much higher than the traditional classification method. The experimental results show that the spectral dimension reduction using a band selection algorithm can effectively improve the accuracy and robustness of the algorithm, and at the same time, make the identification of Raman spectral species more interpretable. The band selection algorithm determines the key band location of blood stain identification, which is also important for the design of a portable Raman system for law enforcement.

Keywords Blood stain; Raman spectrum; Classification model; Band selection

(Received Aug. 20, 2020; accepted Dec. 20, 2020)

《光谱学与光谱分析》期刊社决定采用 ScholarOne Manuscripts 在线投稿审稿系统

《光谱学与光谱分析》期刊社与汤森路透集团签约，自 2010 年 12 月 1 日起《光谱学与光谱分析》决定采用 Thomson Reuters旗下的 ScholarOne Manuscripts 在线投稿审稿系统。

· ScholarOne Manuscripts，该系统不仅能轻松处理稿件，而且能提速科技交流。

· 全球已有 360 多家学会和出版社的 3 800 多种期刊选用了 ScholarOne Manuscripts 系统作为在线投稿、审稿平台，全球拥有超过 1 350 万的注册用户，代表着全球学术期刊在线投审稿的一流水平。

· ScholarOne Manuscripts 与 EndNote，Web of Science 无缝链接和整合；使科研探索、论文评阅和信息传播效率大为提高。

· ScholarOne Manuscripts 是汤森路透科技集团的一个业务部门，拥有丰富的学术期刊业务经验，为学术期刊提供综合管理工作流程系统，使期刊更有效管理投稿、同行评审、加工和发表过程，提高作者心中的专业形象，缩短论文发表时间，削减管理成本，帮助期刊提高科研绩效和实现学术创新。

《光谱学与光谱分析》采用“全球学术期刊首选的在线投稿审稿系统—ScholarOne Manuscripts”，势必对 2010 年 11 月 30 日以前向本刊投稿的作者在查阅稿件信息时，会带来某些不便，在此深表歉意！为了推进本刊的网络化、数字化、国际化进程，以实现与国际先进出版系统对接；为了不断提高期刊质量，加快网络化、数字化建设，加快与国际接轨的进程，希望能得到广大作者、读者们的支持与理解，对您的理解和配合深表感激。这是一件新事物，肯定有不周全、不完善的地方，让我们共同努力，不断改进和完善起来。

《光谱学与光谱分析》期刊社

2010 年 12 月 1 日