第 51 卷 第 5 期/2024 年 3 月/中国激光 研究论文

中国激光

一种基于特征提取的生物气溶胶遥测识别算法研究

杨荣[1,2,3]，董吉辉[1,2,3*]，苏博家[1,2,3]，杨泽后[1,2,3,4]，陈涌[1,2,3,5]，李晓锋[1,2,3,5]，陈春利[1,2,3]，周鼎富[1,2,3]

[1]西南技术物理研究所，四川 成都 610041；

[2]激光雷达与器件技术四川省国防科技重点实验室，四川 成都 610041；

[3]中国兵器工业集团有限公司激光器件技术重点实验室，四川 成都 610041；

[4]北京理工大学物理学院，北京 100081；

[5]北京理工大学光电学院，北京 100081

摘要 荧光激光雷达对气溶胶云团进行远程侦测时，常利用决策树法对云团的荧光光谱信号进行识别。当大气能见度较差或背景辐射较强时，激光雷达的信噪比下降，导致分类识别的准确性明显降低。针对这一问题，提出了一种基于特征提取的决策树分类方法，该方法充分利用荧光光谱信号的信息，具有较强的适用性。首先介绍了生物荧光光谱的特点及传统识别算法和改进识别算法的原理；然后实验测试了6种生物溶液的荧光谱，并通过在这6种生物物质的荧光光谱中增加不同强度的噪声，对两种分类识别算法的性能进行了对比分析。结果表明：所设计的基于特征提取的决策树算法的训练时间基本不随噪声大小改变，当光谱信号的信噪比为10时，对6种生物物质的识别准确率基本达到80%以上；对于两种荧光光谱极其相似的生物，具有较强的区分能力，识别准确性优于传统识别算法；抗噪能力较强，提高了生物气溶胶激光雷达的探测识别能力。

关键词 遥感；激光雷达；激光诱导荧光；机器学习；决策树；生物识别

中图分类号 TN249 **文献标志码** A **DOI**：10.3788/CJL230847

1 引 言

激光诱导荧光雷达基于生物荧光效应，通过远程接收荧光信号来获取相关的生物信息，对于有害生物成分的防区外预警、降低生物袭击带来的毁伤有着重要的作用[1-2]。利用荧光谱识别生物种类是激光雷达遥测中的重要一环，主要算法有决策树、支持向量机和人工神经网络等[3-4]。其中，决策树算法主要是通过选定光谱的一系列特征，对测试集的光谱数据进行训练，得到一棵决策判别树，按照特征对光谱进行一层一层的划分，从而完成识别[5]。

决策树算法的原理简单，易于理解和实现，在荧光谱识别生物中得到了广泛应用。德国航空航天中心在2016年利用决策树算法对355 nm和280 nm激光激发下的荧光谱进行了分类研究，其对不同细菌的鉴别准确率达到了90%[6]。英国曼彻斯特大学在2017年研究了决策树算法的性能，其在两个荧光数据集中的准确率达到了82.8%[7]。波兰军事工业大学光电研究所在2018年利用决策树对48种生物气溶胶进行了实时分类，结果显示，其能正确区分生物气溶胶的种类[8]。

生物荧光光谱与生物所处的生命周期阶段和浓度大小等因素密切相关，其光谱强度的具体数值是波动的[9]。目前大多决策树算法都选用光谱不同波段上的强度作为特征，并未对统计特征进行提取[10-11]，对于同一环境条件下测得的荧光光谱，其识别效果很好，但是一旦光谱受到扰动，所建立的决策树模型可能存在过拟合的情况，导致识别率过低。对于生物气溶胶激光雷达，由于大气状态和背景辐射的不确定性很大，所获取的荧光谱易受扰动，故不适合直接采用决策树算法进行分类识别。

针对该问题，本文提出了一种基于荧光光谱特征提取的决策树识别方法。该方法的特点是利用离散余弦变换(DCT)、中心峰位置、光谱面积建立每张光谱的一系列特征，基于这些特征训练出决策判别树。该方法合理利用了光谱中每个点的信息，减少了过拟合产生的误判，增加了算法在生物识别上的适应性。

2 生物荧光谱的测量与决策树算法简介

2.1 生物荧光谱的测量

利用三维荧光光谱仪，对松花粉、油菜花粉、玫瑰

收稿日期：2023-05-18；**修回日期**：2023-07-09；**录用日期**：2023-08-07；**网络首发日期**：2023-08-15

基金项目：北方激光研究院有限公司青年科技创新项目(K220042-003)

通信作者：*j.h.dong@163.com

花粉、核黄素、辅酶(NADH)、色氨酸等6种典型生物物质的荧光谱进行了测量。设置激发波长为355 nm，积分时间为5 s，截取370～700 nm波段的发射光谱进行荧光光谱分析[12]。

为了减少测量错误，降低实验偶然性的影响，将生物荧光光谱作为训练集和测试集，对每种生物物质重复测量105次，共获得6种生物物质的630张光谱。得到荧光数据后，截取370～700 nm波段的荧光光谱，对其进行归一化处理。105次测量得到的同种物质的荧光谱如图1所示。

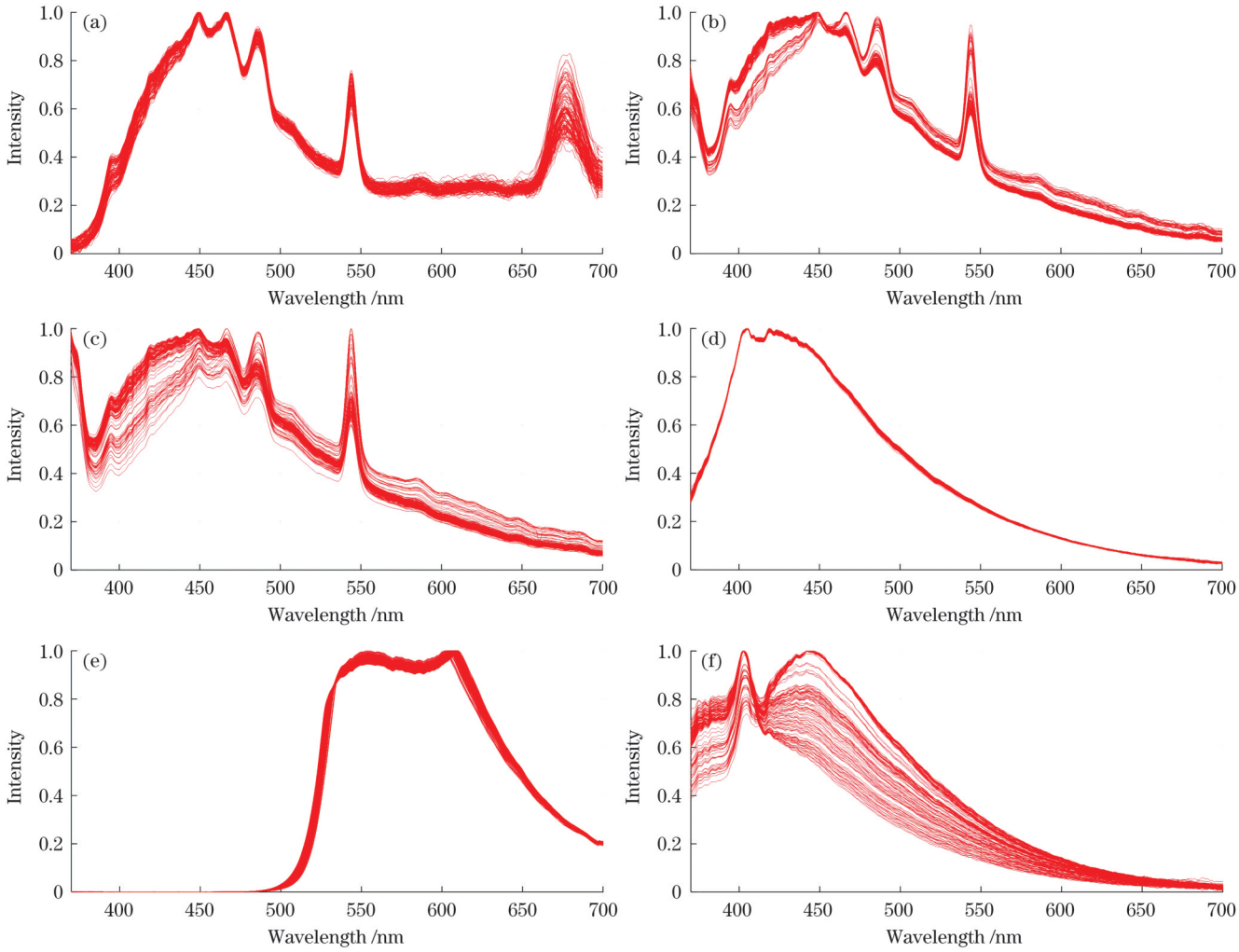

图1　355 nm激发下105次测量得到的各种生物物质的原始荧光光谱图。(a)玫瑰花粉；(b)油菜花粉；(c)松花粉；(d)辅酶；(e)核黄素；(f)色氨酸

Fig. 1　Original fluorescence spectra of various biological substances obtained by 105 measurements under 355 nm excitation. (a) Rose pollen; (b) canola pollen; (c) pine pollen; (d) NADH; (e) riboflavin; (f) tryptophan

由图1可以看出，随着测量次数的增加，除了辅酶和核黄素，其他4种物质的光谱均有变化，但同种物质的波峰位置和基本谱形基本不变，体现了利用光谱识别生物物质的可行性。同种生物物质在不同测量次数下谱形不同的原因，可能是随着时间的变化，溶液中的生物物质发生沉淀，生物物质处于不同生命周期，或者测量仪器本身光功率略微发生变化等，这些情况在实际的激光雷达探测中较为普遍[13-14]。

从6种生物物质荧光光谱可以看出，3种花粉的荧光光谱比较相近，特别是油菜花粉和松花粉，光谱形状极其相似，仅仅在370 nm波段略有区别，该处松花粉的波峰更高。而辅酶、核黄素、色氨酸的荧光光谱形状相差甚远，能与其他物质的荧光光谱区别开。

为了实现对有害生物的提前预警，迫切需要一种可对激光雷达探测到的荧光谱信号进行快速有效识别的算法。目前，荧光光谱的识别算法有很多，其中决策树算法因其过程简单、准确率高，被广泛应用于激光雷达快速预警领域，故本文采用决策树算法。

2.2　决策树分类识别算法

决策树算法利用初始特征A(根节点)对样本的类别进行判断，样本被判断为类别1或者非类别1。在另外一个分支中，利用特征B(非叶节点)对样本的类别

进行判断，样本被判断为类别 2 或者类别 3，每个被判断出来的类别被称为叶节点。利用各个特征对样本进行测试，每一个样本被分类到具体的类别中[15]，如图 2 所示。

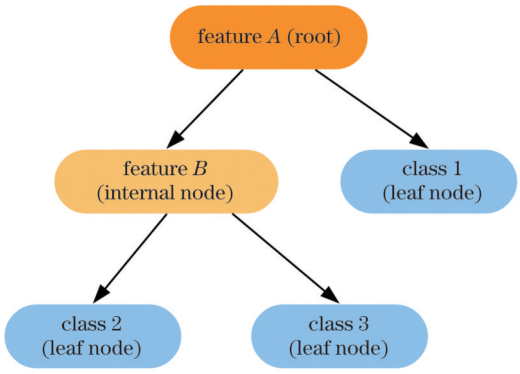

图 2 决策树算法原理图

Fig. 2 Schematic of decision tree algorithm

决策树的训练和建立是决策树分类算法的关键，一个合适的决策树在未知样本出现时能够快速准确地得到最终的决策结果。目前决策树的建立有很多种实现算法，ID3（Iterative Dichotomiser 3）、C4.5（Classifier 4.5）、CART（Classification and Regression Tree）是较早出现的三种方法。ID3 算法的核心是采用信息增益作为特征选取的依据，选择信息增益最大的特征进行树状分裂，其缺点是容易过拟合，只能用来处理离散分布的特征。ID3 算法对可取值数目多的特征具有偏向性，C4.5 算法克服了 ID3 算法的缺点，引入信息增益率作为特征选取的指标，其缺点是耗时长，用的是比二叉树更低效的多叉树，只能用来分类。CART 算法采用基尼系数作为特征选取的依据，没有对数运算，且 CART 算法生成的是二叉树，能减小决策树的规模，提高效率。本文在 ID3 以及 C4.5 算法的基础上，采用改进后的 CART 算法进行训练，具体流程[16]如下。

创建决策树的根节点 A，根节点包含了所有训练集。通过特征划分进行递归，直至叶节点。有以下两种情况可以通过直接递归得到叶节点：第一种，当前节点的所有样本属于同一类别；第二种，候选特征集为空或所有样本的特征值相等。

若该内部节点不会被递归至叶节点，则计算各个候选特征的 Gini 系数，选择最小的 Gini 系数对应的特征作为划分依据，标记为特征 B。Gini 系数的计算过程如下。选取 i 为该特征的划分临界点，将训练集 S 分为 S_1 和 S_2 两部分，那么该特征的 Gini 系数[17]为

$$\mathrm{Gini}(S,i)=\frac{S_1}{S}\left(1-\sum_{S_1}P_j^{\ 2}\right)+\frac{S_2}{S}\left(1-\sum_{S_2}P_j^{\ 2}\right),\quad(1)$$

式中：P_j 是 S_1 或者 S_2 中第 j 类的概率。不断重复上述过程直到生成完整的决策树。

3 决策树算法对生物荧光谱的识别性能分析

3.1 生物气溶胶荧光谱信号模拟

为验证设计的生物荧光光谱识别算法的性能，模拟激光雷达在不同环境下的识别能力，针对建立的生物荧光光谱数据库中的 630 张光谱，增加不同程度的加性噪声，信噪比（S_N）分别为 20、10、5。信噪比取决于纯信号能量（p_s）和纯噪声能量（p_n）的比值[18-19]：

$$S_N=10\lg\left(\frac{p_s}{p_n}\right)。\quad(2)$$

图 3 给出了 6 种生物物质的原始荧光光谱及增加不同噪声后的归一化荧光光谱，可以看到，随着信噪比的降低，噪声明显增加，并且光谱的谱形也会发生改变，比如基线逐渐抬高，谱形变得更宽。

3.2 传统决策树算法的特征选取及识别性能分析

生物气溶胶激光雷达探测到的荧光光谱的光强通常对应几十到几百个不同波段，可将探测到的荧光光谱视为一个离散信号序列 $x(n)$，其中 n 代表第 n 个波段。目前，传统算法将光谱信号各个波段上的强度 $x(n)$ 本身作为备选特征[20]，这种方式意味着一张光谱图具有几十上百个备选特征，而利用决策树进行训练，决策树最终选取用作判别的特征较少，仅用一张光谱图中的少量点信息对生物种类进行判别，导致算法的抗噪能力很弱，损失了大量有用信息。

传统算法的性能随着噪声强度的增加而急剧变弱，当信噪比为 20 以上时算法对各种生物物质的识别准确率维持在 90% 以上，但在生物气溶胶激光雷达的探测下，一般信号的信噪比为 10，此时传统算法的识别准确率大大降低，油菜花粉的识别准确率甚至低于 60%，算法性能显然难以满足激光雷达遥测的要求。

在原始的 630 张光谱中随机抽取 315 张用作训练集，训练出的决策判别树如图 4（a）所示，其中 $x(n)$ 为第 n 个波段的强度，$X(k)$ 为频率阶数为 k 时对应余弦分量的幅值参数。利用训练好的决策树，分别对剩下的原始光谱和加了噪声之后的光谱进行种类判决，以测试算法的性能。

为避免单次训练下准确度过高或者过低的偶然情况，全面评价传统算法的性能特点，必须循环验证 100 次，计算最终的平均准确率（识别正确光谱数与测试光谱数的比值）以及平均运行时间，结果如表 1 所示。

3.3 改良决策树算法的特征选取及识别性能分析

在传统算法的决策树建立过程中会损失大量原始信息，针对该问题，设计一种算法，该算法充分利用荧光光谱每个波段上的点信息，且适应决策树自身特征个数少的特点。通过数学变换，将荧光光谱离散序列信号转换为包含大部分信息的几个特征。比如从人脸

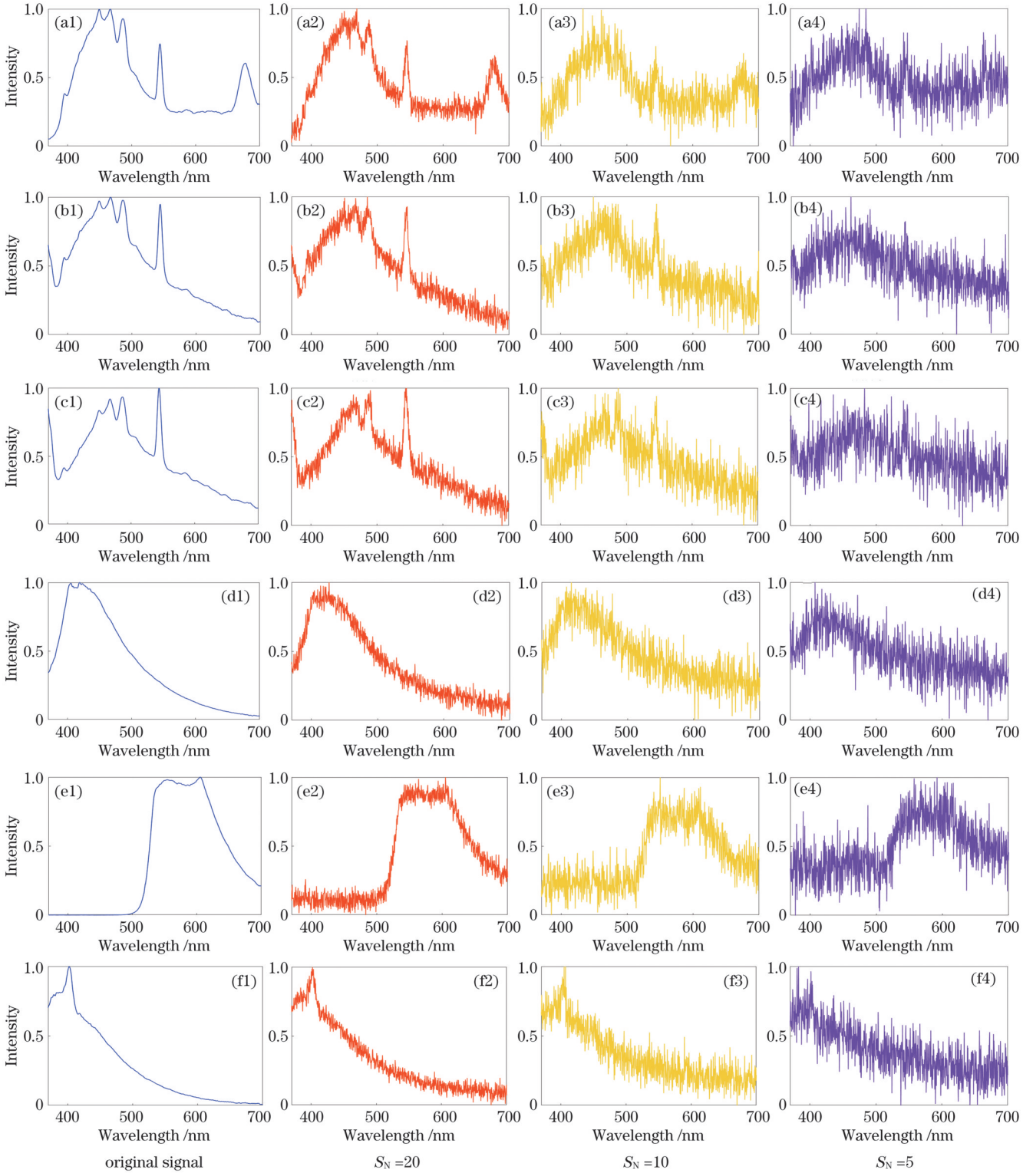

图 3　355 nm 激发下加入不同噪声后各种生物物质的荧光光谱。(a1)～(a4)玫瑰花粉;(b1)～(b4)油菜花粉;(c1)～(c4)松花粉;(d1)～(d4)辅酶;(e1)～(e4)核黄素;(f1)～(f4)色氨酸

Fig. 3　Fluorescence spectra of various biological substances after adding different noises under 355 nm excitation. (a1)-(a4) Rose pollen; (b1)-(b4) canola pollen; (c1)-(c4) pine pollen; (d1)-(d4) NADH; (e1)-(e4) riboflavin; (f1)-(f4) tryptophan

中提取出眼距、肤色等特征,而不是仅仅将人脸图像某个像素点作为特征。

离散余弦变换(DCT)利用少量参数对信号进行表征。在生物气溶胶探测中,不同生物的荧光光谱的波峰位置、归一化后的光谱面积通常是不同的,故利用以上三个方法对特征进行提取。图 5 所示为原始算法利用到的光谱点数和改良后算法利用到的光谱点数对比示意图,传统算法仅用到少量的光谱点信息,而改良

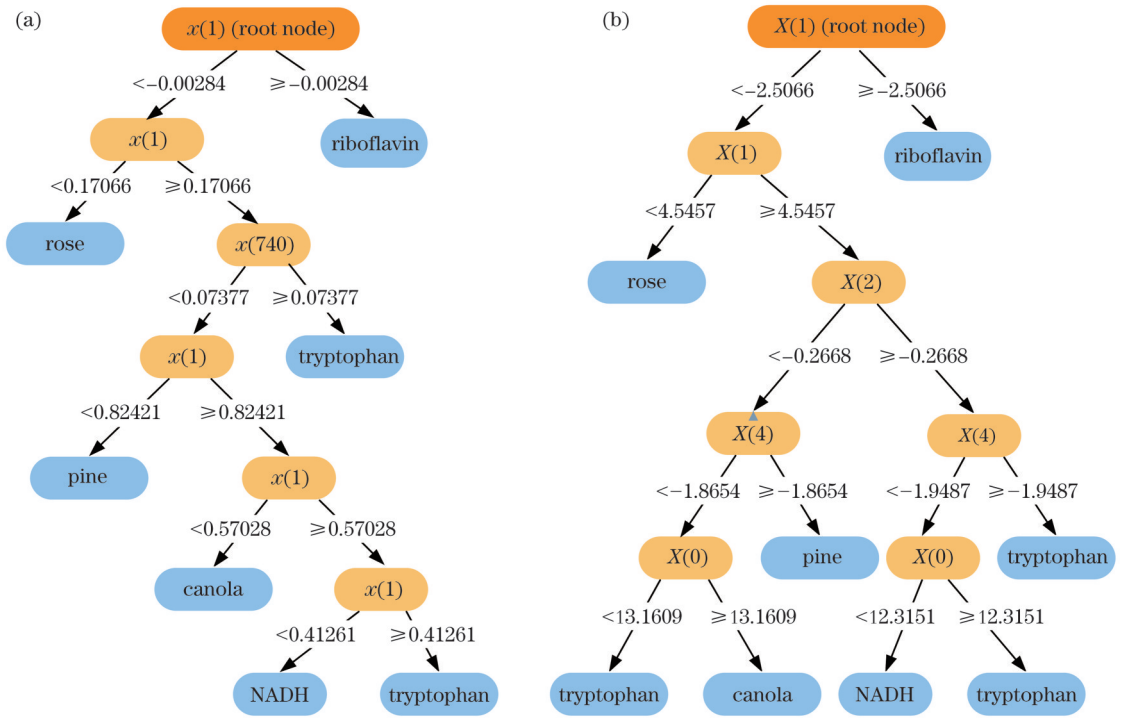

图 4　建立的决策判别树。(a)传统算法的决策判别树;(b)改良后算法的决策判别树

Fig. 4　Established decision discriminant trees. (a) Decision discriminant tree for traditional algorithm; (b) decision discriminant tree for improved algorithm

表 1　传统算法的测试结果

Table 1　Test results of traditional algorithm

Signal-to-noise ratio /dB	Average accuracy /%						Time /ms
	Rose	Canola	Pine	NADH	Riboflavin	Tryptophan	
Original	98.52	95.41	97.20	99.60	99.39	96.00	84.12
20	95.79	95.67	96.57	95.01	96.57	92.53	116.60
10	87.26	59.58	67.63	71.51	96.22	68.09	393.6
5	72.09	36.13	49.39	47.15	83.19	44.15	509.1

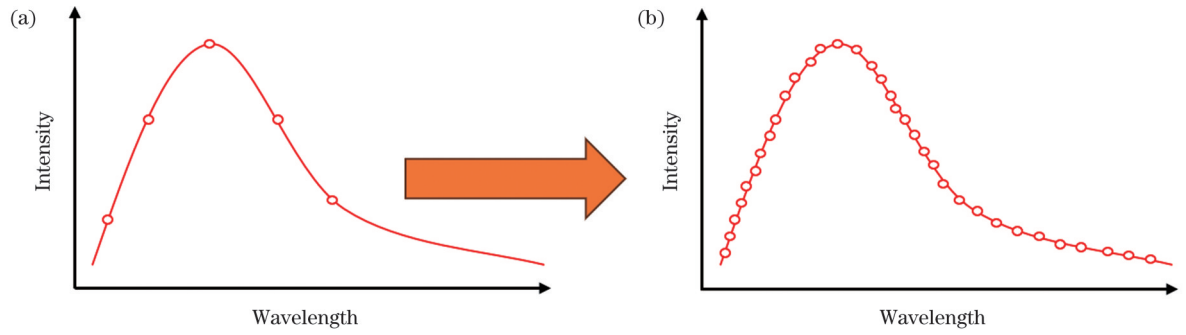

图 5　改良前后算法利用的光谱点信息数量的对比示意图。(a)改良前;(b)改良后

Fig. 5　Comparison diagrams of number of spectral point information used by algorithm before and after improvement. (a) Before improvement; (b) after improvement

后算法几乎用到全部光谱点信息,解决了信息利用不充分的问题。

改良后的特征提取过程主要分为下面三个部分,最终提取到 7 个特征。

3.3.1　DCT

在离散余弦变换中,任何离散信号都能表示为多个具有不同振幅和频率的余弦信号的叠加[21]。数学上共存在 8 种类型的 DCT,最常用的是 DCT-2,公式[22]为

$$x(n)=\sqrt{\frac{1}{N}}X(0)+\sqrt{\frac{2}{N}}\sum_{k=1}^{N-1}X(k)\cos\frac{k(2n+1)\pi}{2N},\tag{3}$$

$$X(k)=\sqrt{\frac{2}{N}}\varepsilon(k)\sum_{n=0}^{N-1}x(n)\cos\frac{k(2n+1)\pi}{2N},\quad k=0,1,\cdots,N-1,\tag{4}$$

式中：$\varepsilon(k)$为系数，在$k=0$时取$\sqrt{\frac{1}{2}}$，$k\neq0$时取1；N为光谱离散信号个数。

离散余弦变换具有能量集中性，即光谱的大部分信息集中在低频段，噪声等信息集中在高频部分。以玫瑰花粉荧光光谱信号序列为例，计算其余弦分量的幅值参数$X(k)$。从图6可以看到，当频率阶数k大于4时，参数$X(k)$很小，所以对于光谱特征，选用DCT后的前5个参数$X(0)$、$X(1)$、$X(2)$、$X(3)$、$X(4)$进行表征。

3.3.2　中心峰

中心峰是一个光谱峰值对应的位置，光谱的中心峰由于噪声的影响，可能存在多个毛刺，影响峰位的判断。这里采用平均平滑的方法来避免这一情况，窗口为50，平滑之后再选取峰值对应位置$n_{\max}$作为第6个特征。

3.3.3　光谱面积

光谱面积也是光谱的一个典型特征，对于离散的光谱信号，光谱面积(σ)的计算公式为

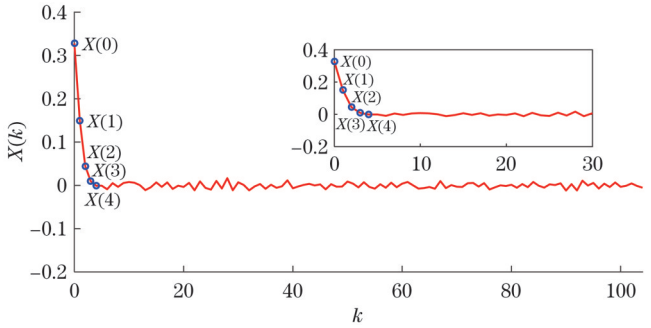

图6　DCT后玫瑰花粉荧光光谱信号的$X(k)$随k的变化

Fig. 6　$X(k)$ versus $k$for fluorescence spectral signal of rose pollen after DCT

$$\sigma=\sum_{n=0}^{N-1}x(n)\cdot\Delta n,\ \Delta n=1,\tag{5}$$

式中：Δn代表离散信号$x(n)$中每两个点的间隔。计算后将光谱面积σ作为第7个特征。

与3.2节同理，在原始的630张光谱中随机抽取315张用作训练集，利用改良后的算法训练出合适的决策树，训练出的决策判别树如图4(b)所示。循环训练测试100次，结果如表2所示。

表2　改良后算法的测试结果

Table 2　Test results of improved algorithm

Signal-to-noise ratio /dB	Average accuracy /%						Time /ms
	Rose	Canola	Pine	NADH	Riboflavin	Tryptophan	
Original	99.98	99.90	98.64	99.34	100.00	96.24	16.56
20	100.00	93.54	99.17	99.12	100.00	91.97	24.13
10	99.90	79.83	88.93	95.10	100.00	84.88	22.15
5	100.00	67.43	74.88	91.13	100.00	73.29	32.98

3.4　两种决策树识别算法的性能对比

对比两者的识别结果，从图7可以明显看出，两种算法在信号信噪比较高时都能对每种生物物质进行较为准确的识别判断(信噪比在20以上时识别率都在90%以上)，但是本文设计的算法明显具有更强的抗噪声性能。观察两种算法的准确率曲线，随着噪声强度的增加，传统算法准确率的下降幅度明显大于本文设计算法。即使在信号信噪比为5时，本文算法的识别准确率基本维持在65%以上，而传统算法的识别准确率大部分已经下降到50%以下，这样是无法满足实际激光雷达探测需求的。另外，从图7还可以看到，玫瑰花粉和核黄素的识别准确率明显高于其他生物物

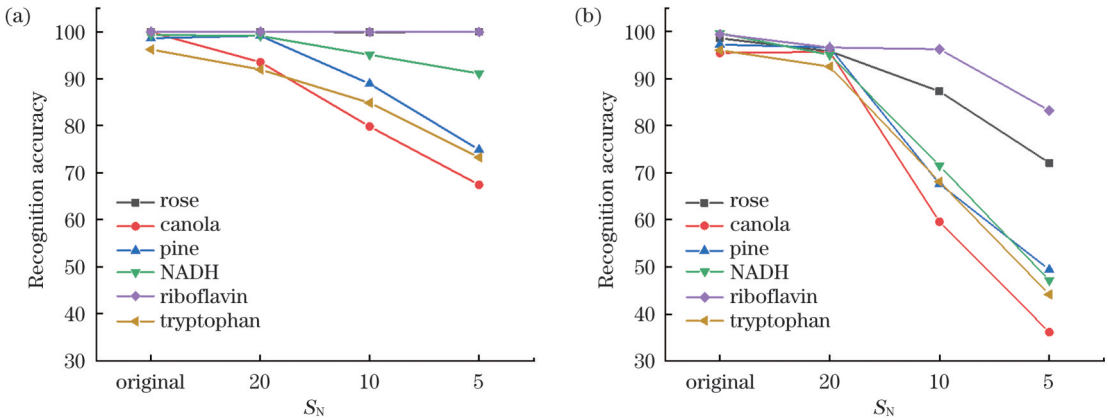

图7　两种算法的识别准确率随信号S_N的变化。(a)改良后算法；(b)传统算法

Fig. 7　Recognition accuracy versus S_N of signal for two algorithms. (a) Improved algorithm; (b) traditional algorithm

质，这是因为它们的荧光光谱和其他4种物质的荧光光谱有着显著差异，均有明显的特征峰（如图3所示，玫瑰花粉明显的特征峰在675 nm左右，核黄素明显的特征峰在550～600 nm处）。

其次本文设计算法的单次训练时间为16～32 ms，远小于传统算法，基本不会随着噪声强度的增大而增加，传统算法的训练时间为84～509 ms，且随着噪声强度的增大而急剧增加。这是因为本文设计算法提取的7个特征量包含光谱的所有信息，无论噪声强度多大，始终只有7个特征，但传统算法的特征量会随着噪声强度的增加而增加。

综上所述，本文设计的改良算法在抗噪声能力上强于传统算法，且训练时间较少。结果显示，本文设计算法完全可行。

4 结　论

提出了一种新型的基于荧光谱统计特征提取的决策树算法，该算法将多个初级特征转换为7个主要高级特征，这7个主要特征涵盖了几乎所有的光谱信息，故算法的抗干扰性优于传统算法。

测试了6种生物物质的荧光光谱，并在每种荧光光谱中分别添加S_N为20、10、5的加性高斯白噪声。使用传统算法和设计算法识别不同的带噪声仿真信号，发现随着信号信噪比的降低，传统算法的识别准确率下降明显，而设计算法的识别准确率保持在较高水平。当S_N=10时，传统算法的识别准确率甚至低于60%，设计算法的识别准确率基本在80%以上，且设计算法的训练速度更快，基本不随噪声变化，各方面性能都优于传统算法。实验结果表明：设计算法提高了识别准确率和训练速度，能够更好地避免误判情况，有助于提升生物气溶胶激光雷达的探测性能。

参 考 文 献

[1] Huffman J A, Perring A E, Savage N J, et al. Real-time sensing of bioaerosols: review and current perspectives[J]. Aerosol Science and Technology, 2020, 54(5): 465-495.

[2] 梁晓峰, 杨泽后, 王顺艳, 等. 基于差分吸收激光雷达有毒有害气体遥测进展[J]. 激光技术, 2021, 45(1): 53-60.
Liang X F, Yang Z H, Wang S Y, et al. Progres in remote sensing of toxic and harmful gas based on differential absorption lidar[J]. Laser Technology, 2021, 45(1): 53-60.

[3] 杨荣, 董吉辉, 杨泽后, 等. 生物气溶胶激光远程侦测技术进展[J]. 激光杂志, 2023, 44(1): 1-7.
Yang R, Dong J H, Yang Z H, et al. Progress in bioaerosol laser remote detection technology[J]. Laser Journal, 2023, 44(1): 1-7.

[4] 陈玉宝, 王箫鹏, 步志超, 等. 超大城市试验气溶胶激光雷达标定及结果分析[J]. 激光技术, 2022, 46(4): 435-443.
Chen Y B, Wang X P, Bu Z C, et al. Calibration and result analysis of aerosol LiDAR in megacity experiment[J]. Laser Technology, 2022, 46(4): 435-443.

[5] Carestia M, Pizzoferrato R, Gelfusa M, et al. Development of a rapid method for the automatic classification of biological agents' fluorescence spectral signatures[J]. Optical Engineering, 2015, 54(11): 114105.

[6] Duschek F, Fellner L, Gebert F, et al. Standoff detection and classification of bacteria by multispectral laser-induced fluorescence [J]. Advanced Optical Technologies, 2017, 6(2): 75-83.

[7] Ruske S, Topping D O, Foot V E, et al. Evaluation of machine learning algorithms for classification of primary biological aerosol using a new UV-LIF spectrometer[J]. Atmospheric Measurement Techniques, 2017, 10(2): 695-708.

[8] Leśkiewicz M, Kaliszewski M, Włodarski M, et al. Improved real-time bio-aerosol classification using artificial neural networks[J]. Atmospheric Measurement Techniques, 2018, 11(11): 6259-6270.

[9] 刘俊秀, 杜彬, 邓玉强, 等. 基于差分-主成分分析-支持向量机的有机化合物太赫兹吸收光谱识别方法[J]. 中国激光, 2019, 46(6): 0614039.
Liu J X, Du B, Deng Y Q, et al. Terahertz-spectral identification of organic compounds based on differential PCA-SVM method[J]. Chinese Journal of Lasers, 2019, 46(6): 0614039.

[10] Gebert F, Kraus M, Fellner L, et al. Novel standoff detection system for the classification of chemical and biological hazardous substances combining temporal and spectral laser-induced fluorescence techniques[J]. The European Physical Journal Plus, 2018, 133(7): 269.

[11] Narlagiri L M, Bharati M S S, Beeram R, et al. Recent trends in laser-based standoff detection of hazardous molecules[J]. TrAC Trends in Analytical Chemistry, 2022, 153: 116645.

[12] Pan Y L. Detection and characterization of biological and other organic-carbon aerosol particles in atmosphere using fluorescence [J]. Journal of Quantitative Spectroscopy and Radiative Transfer, 2015, 150: 12-35.

[13] 王其, 曾万聃, 夏志平, 等. 基于随机森林算法的食源性致病菌拉曼光谱识别[J]. 中国激光, 2021, 48(3): 0311002.
Wang Q, Zeng W D, Xia Z P, et al. Recognition of food-borne pathogenic bacteria by Raman spectroscopy based on random forest algorithm[J]. Chinese Journal of Lasers, 2021, 48(3): 0311002.

[14] 余晓娅, 张玉钧, 殷高方, 等. 基于偏最小二乘回归的藻类荧光光谱特征波长选取[J]. 光学学报, 2014, 34(9): 0930002.
Yu X Y, Zhang Y J, Yin G F, et al. Feature wavelength selection of phytoplankton fluorescence spectra based on partial least squares [J]. Acta Optica Sinica, 2014, 34(9): 0930002.

[15] 季桂树, 陈沛玲, 宋航. 决策树分类算法研究综述[J]. 科技广场, 2007(1): 9-12.
Ji G S, Chen P L, Song H. Study the survey into the decision tree classification algorithms rule[J]. Science Mosaic, 2007(1): 9-12.

[16] 赖春廷. 决策树分类算法研究[J]. 信息与电脑(理论版), 2020, 32(14): 59-62.
Lai C T. Research on decision tree classification algorithm[J]. China Computer & Communication, 2020, 32(14): 59-62.

[17] 王珏, 周志华, 周傲英. 机器学习及其应用[M]. 北京: 清华大学出版社, 2006.
Wang J, Zhou Z H, Zhou A Y. Machine learning and its application[M]. Beijing: Tsinghua University Press, 2006.

[18] 丁红波, 王珍珠, 刘东. 激光雷达信号去噪方法的对比研究[J]. 光学学报, 2021, 41(24): 2401001.
Ding H B, Wang Z Z, Liu D. Comparison of de-noising methods of LiDAR signal[J]. Acta Optica Sinica, 2021, 41(24): 2401001.

[19] Hu M H, Mao J D, Li J, et al. A novel lidar signal denoising method based on convolutional autoencoding deep learning neural network[J]. Atmosphere, 2021, 12(11): 1403.

[20] Fellner L, Kraus M, Gebert F, et al. Multispectral LIF-based standoff detection system for the classification of CBE hazards by spectral and temporal features[J]. Sensors, 2020, 20(9): 2524.

[21] 冈萨雷斯. 数字图像处理[M]. 北京: 电子工业出版社, 2011.
Rafael C G. Digital image processing[M]. Beijing: Publishing House of Electronics Industry, 2011.

[22] 程佩青. 数字信号处理教程: MATLAB版[M]. 5版. 北京: 清华大学出版社, 2017.
Cheng P Q. Digital signal processing course: MATLAB version [M]. 5th ed. Beijing: Tsinghua University Press, 2017.

Feature Extraction-Based Bioaerosol Telemetry Identification Algorithm

Yang Rong[1,2,3], Dong Jihui[1,2,3*], Su Bojia[1,2,3], Yang Zhehou[1,2,3,4], Chen Yong[1,2,3,5], Li Xiaofeng[1,2,3,5], Chen Chunli[1,2,3], Zhou Dingfu[1,2,3]

[1]*Southwest Institute of Technical Physics, Chengdu* 610041, *Sichuan, China*;

[2]*Sichuan Provincial Key Laboratory of National Defense Science and Technology of LiDAR and Device Technology, Chengdu* 610041, *Sichuan, China*;

[3]*Key Laboratory of Laser Device Technology, China North Industries Group Corporation Limited, Chengdu* 610041, *Sichuan, China*;

[4]*College of Physics, Beijing Institute of Technology, Beijing* 100081, *China*;

[5]*College of Optoelectronics, Beijing Institute of Technology, Beijing* 100081, *China*

Abstract

Objective In the remote detection of bioaerosol clouds by fluorescence lidar, the decision tree method is often used to identify the fluorescence spectral signals of the clouds. The conventional decision tree algorithm selects the intensity values of the echo signals at different wavebands as features rather than extracting the statistical features of the echo signals, thereby effectively recognizing the fluorescence spectra measured under the same environmental conditions. However, in bioaerosol LiDAR, the acquired fluorescence spectra are highly variable because of the great uncertainty of the atmospheric state and background radiation, such that when the signal-to-noise ratio of LiDAR decreases, the previously established decision tree model may be overfitted, resulting in low recognition accuracy. In this study, the conventional algorithm is improved to increase the noise resistance of recognition and make the algorithm applicable to the field of LiDAR detection of bioaerosols.

Methods In this study, fluorescence spectral signals of six biomaterials are first tested under laboratory conditions. Different Gaussian white noises with different intensity values are added to the fluorescence spectrum of each material to simulate the actual echo signals detected by bioaerosol LiDAR. Subsequently, the fluorescence spectra and recognition algorithms are analyzed mechanistically, and a decision tree recognition algorithm based on statistical feature extraction is designed, primarily based on discrete cosine transform (DCT), central peak position, and normalized spectral area. Finally, the performance of the two recognition algorithms is examined with simulated LIDAR signals under different noise intensity values. The two algorithms are used to train the spectra of the training set to form their respective decision trees, concurrently recording the training time. The decision trees are used to discriminate the test set, whereby the accuracy is calculated to analyze the actual detection ability of the algorithms before and after the improvement.

Results and Discussions Both algorithms accurately recognize each biomass when the signal-to-noise ratio (SNR) of the signal is high. The recognition rate is above 90% when the SNR is above 20. However, the performance of the traditional algorithm dramatically weakens with an increase in noise. In the detection of bioaerosol LiDAR, the SNR is 10, leading to greatly reduced recognition accuracies of the traditional algorithms. The recognition accuracy of rapeseed pollen is lower than 60%. When the SNR is 5, the recognition accuracies are even lower than 50% for the four kinds of substances, clearly making it difficult to support the performance of the algorithms to meet the requirements of LiDAR telemetry. The improved algorithm maintains a recognition accuracy of above 65% even when the SNR is 5, and the recognition accuracy is above 80% when SNR is 10. Second, the training time of the algorithm designed in this study is 16–32 ms, which is much smaller than that of the traditional algorithm. This training time does not increase with the noise intensity, whereas the training time of the traditional algorithm, which is 84–509 ms, sharply increases with the noise intensity.

Conclusions To solve the problem of efficient recognition of biofluorescence spectra by bioaerosol LiDAR, this study designs a novel decision tree algorithm based on statistical feature extraction of fluorescence spectra, by transforming the original primary multiple features into seven main high-level features through DCT, searching for the central wavelength, and calculating the spectral area, which covers almost all the spectral information. The proposed algorithm is faster to train and more noise-resistant, outperforming the traditional algorithm in all aspects. The results show that the decision tree algorithm based on feature extraction improves recognition accuracy and training speed, thereby averting misclassification and enhancing the detection performance of bioaerosol LiDAR.

Key words remote sensing; LiDAR; laser induced fluorescence; machine learning; decision tree; biometrics