激光写光电子学进展

基于密集连接网络模型的致病菌拉曼光谱分类

杨勇^{1,2},董浩^{1,2},桑瑶烁^{1,2},李志刚^{1,2},张龙^{1,2},王玲¹,王澍^{1,2*} ¹中国科学院合肥物质科学研究院安徽光学精密机械研究所,安徽 合肥 230031; ²中国科学技术大学研究生院科学岛分院,安徽 合肥 230031

摘要 细菌拉曼光谱信号弱、相似度高且易被噪声干扰,使用传统机器学习方法对其分类时必须进行繁杂的光谱预处理,效率低下。为提高细菌拉曼光谱分类的准确率和效率,提出了一种基于密集连接的一维卷积神经网络模型 Ramannet,无需额外的光谱预处理就能有效完成光谱分类。实验结果表明,Raman-net对 Bacteria-ID公开数据集中 30 种细菌低 信噪比拉曼光谱的分类准确率为 84.26%,显著高于传统机器学习方法及对比方法。对于碳青霉烯类抗生素敏感和耐药的 2 种肺炎克雷伯菌表面增强拉曼光谱,Raman-net取得了 99.16% 的分类准确率。这表明对于细菌的普通拉曼光谱和表面增强拉曼光谱,Raman-net 取得了 99.16% 的分类准确率。这表明对于细菌的普通拉曼光谱和表面增强拉曼光谱,Raman-net 无需光谱预处理就能取得较好的分类效果,为致病菌的拉曼光谱鉴定提供了一种快速有效的方法。

关键词 光谱学; 拉曼光谱; 光谱鉴别; 机器学习; 致病菌 中图分类号 TP391 **文献标志码** A

DOI: 10.3788/LOP213226

Raman Spectral Classification of Pathogenic Bacteria Based on Dense Connection Network Model

Yang Yong^{1,2}, Dong Hao^{1,2}, Sang Yaoshuo^{1,2}, Li Zhigang^{1,2}, Zhang Long^{1,2}, Wang Ling¹, Wang Shu^{1,2*}

¹Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China;

²Science Island Branch of Graduate School, University of Science and Technology of China, Hefei 230031, Anhui, China

Abstract Bacterial Raman spectrum is characterized by a weak signal, high similarity, and susceptibility to noise. Its classification using traditional machine learning approaches requires complex spectral preprocessing, and the efficiency is low. In this study, to enhance the accuracy and efficiency of bacterial Raman spectral classification, a one-dimensional convolutional neural network model Raman-net based on dense connection is suggested, which could efficiently complete spectral classification without additional spectral preprocessing. The experimental findings demonstrate that the classification accuracy of Raman-net for 30 bacterial low-signal-to-noise ratios Raman spectra in the Bacteria-ID public data set is 84.26%, which is substantially higher than that of traditional machine learning approaches and comparison approaches. Raman-net attained a classification accuracy of 99.16% for surface-enhanced Raman spectroscopy of 2 Klebsiella pneumoniae susceptible and resistant to carbapenems. This demonstrates that Raman-net can attain remarkable classification findings for ordinary Raman spectroscopy and surface-improved Raman spectroscopy of pathogenic bacteria.

Key words spectroscopy; Raman spectroscopy; spectral discrimination; machine learning; pathogenic bacteria

收稿日期: 2021-11-14; 修回日期: 2021-12-15; 录用日期: 2021-12-21; 网络首发日期: 2022-01-05

通信作者: *228690590@qq.com

研究论文

基金项目:国家自然科学基金面上项目(82072380)、安徽省自然科学基金青年项目(1908085QB85)、中国科学院合肥物质科学研究院医疗器械监管专项(YZJJ2021-J-YQ4)

1引言

细菌感染严重危害人类身体健康,但目前临床上 常用的检测手段需要进行细菌培养,操作繁琐且耗时 较长,往往会耽误最佳治疗时机¹¹。因此,开发一种快 速准确的细菌识别方法对及时有效治疗细菌感染具有 重要意义。

拉曼光谱可以提供细菌所含不同生物分子的指纹 图谱信息,如蛋白质、DNA和脂质^[2]。每种细菌都有 其特定的指纹图谱信息,因此,可将拉曼光谱用于细菌 种类鉴定11。但细菌的拉曼信号弱、易被噪声干扰且 同一菌种不同亚型的拉曼光谱差异极小、难以区分[3]。 细菌拉曼光谱分类中应用广泛的是传统机器学习方 法^[4], 如: Kloβ等^[5]用支持向量机(SVM)成功实现了 11种会造成尿路感染细菌的分类;史如晋等^[6]构建的 集成学习模型实现了对大肠杆菌和布鲁氏菌的快速鉴 定;Yan等^[7]结合核主成分分析和决策树方法实现了 23种食源性致病菌的分类。传统机器学习方法对细 菌的分类效果主要取决于光谱预处理策略,且只能采 用尝试法寻找最佳预处理策略,效率不高[8]。卷积神 经网络(CNN)模型具有强大的特征自学习能力,可根 据数据集自动提取最有利于任务的特征^[9]。如:Liu 等^[10]将CNN模型用于矿物的拉曼光谱分类,结果表 明,CNN模型的分类效果优于常见的传统机器学习算 法;Huang 等^[11]使用一维卷积网络模型成功实现了 20种动物血液拉曼光谱的分类;Ho等^[12]设计出一个 基于残差的一维卷积网络模型,实现了30种常见致病 菌拉曼光谱的分类。

本文构建了一种基于密集连接的一维卷积网络模型 Raman-net,无需额外的光谱预处理就能有效对细菌拉曼光谱进行分类。在Bacteria-ID数据集上的实验结果表明,Raman-net的分类准确率远高于传统机器学习方法且优于 Ho 等^[12]提出的残差网络模型。此外,对细菌表面增强拉曼光谱 Raman-net 也能得到较好的分类效果。

2 算法原理

2.1 Raman-net模型

2.1.1 Raman-net 原理

何凯明等^[13]提出的残差网络(ResNet)在普通 CNN模型中引入快捷连接(Shortcut connection),缓解 了训练深层神经网络模型过程中出现的梯度消失和退 化问题(Degradation problem),使深层的网络模型更 易优化,从而通过增加网络模型的层数增强模型的表 达能力,改善模型性能。黄高等^[14]受 ResNet的启发, 在普通卷积网络模型中引入更密集的快捷连接,提出 了密集连接网络(DenseNet)。该模型一经提出就在 MNIST、CIFAR-10和CIFAR-100等图片数据集上取 得了良好的分类性能。DenseNet密集连接块内每个 卷积层的特征图都和前面所有层的特征图在通道维度 上拼接在一起,作为下一网络层的输入,如图1所示。 密集连接块可表示为

$$x_L = H_L([x_0, \cdots, x_{L-1}]),$$
 (1)

式中:[x₀,…,x_{L-1}]为第0层到第L-1层特征图的拼 接;H_L为第L层对输入特征图的操作。DenseNet通过 密集连接对特征图进行复用,既提高了特征图的利用 率,又加强了模型各层间信息的有效传递。尽管拥有 更多的快捷连接,但DenseNet的待学习参数量少于 ResNet,使DenseNet能用更低的硬件条件取得更好的 模型效果。



图 1 密集连接块的结构 Fig. 1 Structure of the dense connection block

受 DenseNet 启发, Raman-net 将密集连接引入一 维 CNN, 提高了模型对拉曼光谱特征图的利用率以及 模型的准确率和鲁棒性。密集连接块内每一层仅向后 层传递少量特征图, 使 Raman-net 的待学习参数量少 于普通 CNN, 给模型的训练和部署带来便利。

2.1.2 Raman-net模型的结构

Raman-net 是一个密集连接的一维 CNN 模型,由 一个初始卷积层(Conv)、4个密集连接块以及完全连接 层等组成。Raman-net 密集连接块的主体结构如图 2 所



图 2 Raman-net 的密集连接块 Fig. 2 Dense connection block of the Raman-net

研究论文

示。初始卷积层有64个卷积核,卷积核尺寸为1×5, 步长为2,使用边缘填充。随后是一个1×3的最大池 化层,步长为2,使用边缘填充。每个基本块由一个 1×1卷积层和一个1×3卷积层组成,步长均为1,使用 边缘填充。4个密集连接块分别由1、2、10、16个基本 块组成,每个密集连接块内的基本块都使用快捷连接 向其之后的基本块传递3个特征图,每两个密集连接 块之间都有一个转换层,用于将上一密集连接块输出 的特征图通道数压缩后传入下一密集连接块。转换层 由步长为1的1×1卷积层接步长为2的1×2平均池化 层组成。最后,由一个1×7平均池化层将特征图压缩 后输入全连接(FC)层得出分类结果,模型最终输出细 菌的种类为C.glabrata,如图3所示。





使用预激活机制,即在模型所有卷积层之前依次 设置批归一化(BN)层和线性整流单元(ReLU)激活 函数。ReLU激活函数可表示为

$$f(x) = \begin{cases} x, x \ge 0\\ 0, x < 0^{\circ} \end{cases}$$
(2)

BN的主要作用是加快模型收敛速度,同时具有 一定的正则化效果。模型训练过程中,采用交叉熵作 为损失函数,可表示为

$$L(x_{i}, y_{i}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log p_{i,k}, \qquad (3)$$

式中:x_i为第i个光谱样本;y_i为第i个光谱样本的标

签; y_{i,k}为将第 i 个样本预测成类别 k; p_{i,k}为将第 i 个样本预测成类别 k 的 概 率; N 为 样 本 总 数; K 为 类 别 总 数。

2.2 对比算法

SVM 是一种线性二分类模型,其目标是在样本线 性可分情况下从样本空间中寻找一个最大间隔分离超 平面完成数据二分类^[15]。通过引入核函数将非线性可 分的数据映射到高维空间中,进而实现对非线性可分 数据的分类,再通过一对一或一对多策略实现 SVM 的多分类模型。

随机森林(RF)是一种优秀的并行集成学习方法, 是许多决策树的集成。对集成模型而言,基学习器间 的差异性越大,模型的效果就越好。由于引入了随机 抽样和随机选取划分属性两个随机步骤,增大了 RF 模型中各决策树间的差异性,进而提升了模型的性能。 RF模型不过度依赖特定样本和属性,具有较强的抗干 扰能力和泛化能力^[16]。

*K*近邻(KNN)算法的思想是先计算待分类样本 到其余各样本的距离(一般采用 Minkowsky 距离或 Euclidean 距离),然后选出距离待分类样本最近的*K* 个样本,最后统计*K*个样本的类别,将待分类样本划分 为*K*个样本中数量最多的类别。KNN算法一般适用 于样本容量较大的数据集,对于样本容量较小的数据 集分类效果不佳^[17]。

3 研究对象

3.1 普通拉曼光谱数据集

Ho等^[12]公开的Bacteria-ID细菌拉曼光谱数据集 包含30种常见致病菌的78500条拉曼光谱,光谱范围 为381.98~1792.4 cm⁻¹,共1000个波段,实验中使用 了该数据集的reference、finetune和test子集。其中: reference子集包含了60000条拉曼光谱数据,30种致 病菌各2000条; finetune和test子集各包括3000条拉 曼光谱数据,30种致病菌各100条。所有拉曼光谱数 据均经过归一化处理并去除基线。

3.2 表面增强拉曼光谱数据集

3.2.1 实验样本

细菌收集自徐州医科大学附属医院检验科,包含 耐碳青霉烯类肺炎克雷伯菌(CRKP)和对碳青霉烯类 敏感的肺炎克雷伯菌(CSKP)各10株。分离株在哥伦 比亚血平板培养基上37℃过夜培养,然后用无菌去离 子水制备浓度约为10⁷ mL⁻¹的细菌溶液。

3.2.2 实验器材及过程

根据 Lee 等^[18]报道的柠檬酸三钠还原法制备纳米 银颗粒(质量浓度为0.1 mg/mL,颗粒粒径为40~ 60 nm)作为增强基底,将10 μL 菌液样品与10 μL 银纳 米颗粒均匀混合在 EP 管中,孵育10 min 后取4 μL 混 合物滴在单晶硅片上作为测量样品。测量前以硅片在

研究论文

 520.7 cm^{-1} 的峰为基准,对便携式拉曼光谱仪(B & W TEK, i-Raman Plus BWS465-785H)进行校准。仪器 设置:用785 nm 氦氖激光作为激发光源,激光功率为 14 mW;20倍长聚焦物镜,积分时间为4 s,积分3次,测 量范围为400~1800 cm⁻¹。测量中每个菌株制备10份 样品,每份样品随机选择6个位点进行检测,共获得 2种肺炎克雷伯菌各600条表面增强拉曼光谱。含有 碳青霉烯酶是CRKP具备耐药性的最主要原因^[19]。 图 4 为 CRKP 和 CSKP 的拉曼光谱图。两种细菌拉曼 光谱的主要差异:CSKP 在拉曼位移 1537、1029、914、 789 cm⁻¹ 处均比 CRKP 多一个子峰,在 1220 cm⁻¹ 处 CRKP的峰比CSKP更平缓。可以发现:在1537 cm⁻¹附 近的峰被认为是C—H₂键的变形^[20];在1220 cm⁻¹附近 的峰被认为由核酸、蛋白质和脂质引起^[21];在1029 cm⁻¹ 附近的峰被认为由苯丙氨酸引起^[22]:在914 cm⁻¹附近的 峰被认为由C-C键引起^[23];在789 cm⁻¹附近的峰被认 为由核酸引起^[21]。





4 实验结果与分析

4.1 普通拉曼光谱数据集实验结果与分析

首先,用 Raman-net 在 reference 数据集上进行训 练得到预训练模型。然后,将预训练模型在 finetune 数 据集上进行微调整。最后,在 test 数据集上测试模型 的泛化效果,得到 Raman-net 的分类准确率。反复尝 试后设置的模型参数:1)用非线性函数 ReLU 作为激 活函数;2)用 dropout 技术及早停策略降低模型的过 拟合程度,drop rate 为0.1;3)批处理大小为16;4)用 Adam 作为优化器,学习速率设置为0.001,betas 设置 为(0.5,0.999);5)采用5折交叉验证方式进行模型 训练。

SVM、RF和KNN模型均使用finetune数据集作 为训练集,test数据集作为测试集,使用网格搜索策略 寻找最佳参数组合,采用5折交叉验证方式训练模型。 为避免偶然性干扰,所有实验均重复5次,取5次分类

第 60 卷 第 1 期/2023 年 1 月/激光与光电子学进展

结果的平均值作为最终分类准确率,结果如表1所示。可以发现:相比SVM、RF和KNN,Raman-net的准确率分别提高了5.13个百分点、20.46个百分点和47.63个百分点;相比Ho等^[12]提出的一维CNN模型,Raman-net的准确率提高了2.26个百分点。

衣 1 小凹候望对音迪拉叟兀盾的刀尖准如	表1	不同模型对普	通拉曼州	6 谱的分	▶类准确	率
----------------------	----	--------	------	-------	------	---

Table 1 Classification accuracy of ordinary Raman spectra by	`able 1	Classification	accuracy of	of ordinary	Raman	spectra b	у
--	---------	----------------	-------------	-------------	-------	-----------	---

	different models	unit: ½	
Model	Accuracy		
Raman-net	84.26		
SVM	79.13		
RF	63.80		
KNN	36.63		

4.2 表面增强拉曼光谱数据集实验结果与分析

为研究数据集大小对各模型分类准确率的影响, 依次从肺炎克雷伯菌表面增强拉曼光谱数据集中取每 种细菌各100、200、300、400、500、600条光谱作为数据 集,所有实验均将数据集随机划分80%作为训练集, 20%作为测试集。无需任何光谱预处理流程,将 Raman-net的预训练模型在训练集上进行训练,在测 试集上测试模型的分类效果。SVM、RF和 KNN 模型 均在训练集上训练,在测试集上得出分类准确率,使用 网格搜索策略寻找最佳参数组合,采用5折交叉验证 方式训练模型。为避免偶然性干扰,所有实验均重复 5次,取5次分类结果的平均值作为最终分类准确率, 结果如表2所示。可以发现,随着光谱数量的增加,各 模型的准确率虽有小幅波动,但整体保持上升趋势,而 Raman-net模型在不同数据量下均取得最高的分类准 确率。尽管CRKP和CSKP的表面增强拉曼光谱极其 相似且存在增强材料特征峰的干扰,Raman-net的分 类准确率仍可达99.16%,显著高于传统机器学习 方法。

表2 不同模型对表面增强拉曼光谱的分类准确率

Table 2	Classification accuracy of surface-enhanced	Raman	
	spectra by different models	unit	%

	- I			
Number of spectra	Raman-net	SVM	RF	KNN
200	75.00	72.50	65.00	72.50
400	90.00	87.50	70.00	77.50
600	96.67	86.67	80.83	84.16
800	97.50	88.13	80.00	83.13
1000	98.00	89.50	83.50	85.50
1200	99.16	89.58	85.42	87.92

5 结 论

Raman-net无需光谱预处理就可有效分类细菌的 拉曼光谱,对30种致病菌的拉曼光谱分类的准确率为

研究论文

84.26%, 优于 SVM(79.19%)、RF(63.80%)和 KNN (36.63%)以及一维 CNN 模型。对于两种极其相似的 肺炎克雷伯菌的表面增强拉曼光谱, Raman-net 的识别 准确率可达 99.16%, 显著高于传统机器学习方法。综 上所述, Raman-net 对细菌的普通拉曼光谱和表面增强 拉曼光谱均具有优秀的分类性能且无需光谱预处理, 为细菌的拉曼光谱分类提供了一种快速有效的手段。

参考文献

- 杨凯,李弘哲,朱永官,等.基于拉曼光谱的快速细菌 诊断[J].光散射学报,2019,31(4):336-345.
 Yang K, Li H Z, Zhu Y G, et al. Rapid bacterial diagnosis by Raman spectroscopy[J]. The Journal of Light Scattering, 2019, 31(4): 336-345.
- [2] 依丽努尔·木合塔尔, 沈爱国, 胡继明. 细菌的拉曼光谱 研究进展[J]. 光散射学报, 2015, 27(2): 110-118.
 Iinur M Shen A G, Hu J M. Advances of Raman spectroscopic studies on bacteria[J]. The Journal of Light Scattering, 2015, 27(2): 110-118.
- [3] 高鹏亚,苏英会,孙晖,等.显微共聚焦拉曼技术在细菌分类鉴定中的应用[J].疾病监测,2021,36(1):74-81.
 Gao P Y, Su Y H, Sun H, et al. Application of micro confocal Raman technique in classification and identification of bacteria[J]. Disease Surveillance, 2021, 36(1):74-81.
- [4] Lussier F, Thibault V, Charron B, et al. Deep learning and artificial intelligence methods for Raman and surfaceenhanced Raman scattering[J]. TrAC Trends in Analytical Chemistry, 2020, 124: 115796.
- [5] Kloβ S, Kampe B, Sachse S, et al. Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study[J]. Analytical Chemistry, 2013, 85(20): 9610-9616.
- [6] 史如晋, 夏钒曾, 曾万聃, 等. 基于 PCA-Stacking 模型的食源性致病菌拉曼光谱识别[J]. 激光与光电子学进展, 2019, 56(4): 043003.
 Shi R J, Xia F Z, Zeng W D, et al. Raman spectroscopic classification of foodborne pathogenic bacteria based on PCA-Stacking model[J]. Laser & Optoelectronics Progress, 2019, 56(4): 043003.
- [7] Yan S S, Wang S Y, Qiu J X, et al. Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level [J]. Talanta, 2021, 226: 122195.
- [8] Engel J, Gerretzen J, Szymańska E, et al. Breaking with trends in pre-processing? [J]. TrAC Trends in Analytical Chemistry, 2013, 50: 96-106.
- [9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015: 1-9.
- [10] Liu J C, Osadchy M, Ashton L, et al. Deep convolutional neural networks for Raman spectrum recognition: a unified solution[J]. The Analyst, 2017, 142

第 60 卷 第 1 期/2023 年 1 月/激光与光电子学进展

(21): 4067-4074.

- [11] Huang S, Wang P, Tian Y B, et al. Blood species identification based on deep learning analysis of Raman spectra[J]. Biomedical Optics Express, 2019, 10(12): 6129-6144.
- [12] Ho C S, Jean N, Hogan C A, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning[J]. Nature Communications, 2019, 10(1): 4927.
- [13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.
- [14] Huang G, Liu Z, Van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.
- [15] Xu Y, Zomer S, Brereton R G. Support vector machines: a recent method for classification in chemometrics[J]. Critical Reviews in Analytical Chemistry, 2006, 36(3/4): 177-188.
- [16] Gromski P S, Muhamadali H, Ellis D I, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis-a marriage of convenience or a shotgun wedding[J]. Analytica Chimica Acta, 2015, 879: 10-23.
- [17] Morais C L M, Lima K M G, Singh M, et al. Tutorial: multivariate classification for vibrational spectroscopy in biological samples[J]. Nature Protocols, 2020, 15(7): 2143-2162.
- [18] Lee P C, Meisel D. Adsorption and surface-enhanced Raman of dyes on silver and gold sols[J]. The Journal of Physical Chemistry, 1982, 86(17): 3391-3395.
- [19] 崔娟,王佳,姚慧生,等.2006—2010年儿童肺炎支原体感染流行病学分析[J].中国实用儿科杂志,2013,28
 (6):446-448.
 Cui J, Wang J, Yao H S, et al. Epidemiologic analysis of MP infection in children from 2006 to 2010[J]. Chinese
- [20] Maquelin K, Kirschner C, Choo-Smith L P, et al. Identification of medically relevant microorganisms by vibrational spectroscopy[J]. Journal of Microbiological Methods, 2002, 51(3): 255-271.

Journal of Practical Pediatrics, 2013, 28(6): 446-448.

- [21] Notingher I, Hench L L. Raman microspectroscopy: a noninvasive tool for studies of individual living cells in vitro[J]. Expert Review of Medical Devices, 2006, 3(2): 215-234.
- [22] Fan C, Hu Z Q, Mustapha A, et al. Rapid detection of food- and waterborne bacteria using surface-enhanced Raman spectroscopy coupled with silver nanosubstrates
 [J]. Applied Microbiology and Biotechnology, 2011, 92(5): 1053-1061.
- [23] Devitt G, Howard K, Mudher A, et al. Raman spectroscopy: an emerging tool in neurodegenerative disease research and diagnosis[J]. ACS Chemical Neuroscience, 2018, 9(3): 404-420.