面向生物医学检测的拉曼光谱图像机器学习算法研究

于铠铭[1a]，包晓栋[1b]，李备[2,3]，洪喜[2]，刘景鑫[1c]

1. 吉林大学中日联谊医院 a. 手外科；b. 医学影像工程中心；c. 放射科，吉林 长春 130033；2. 长春长光辰英生物科学仪器有限公司，吉林 长春 130033；3. 中国科学院 长春光学精密机械与物理研究所，吉林 长春 130033

[摘 要] **目的** 探讨拉曼光谱图像进行快速生物医学检测的机器学习处理分析算法，为建立便捷快速的乙肝创新检测方法提供参考。**方法** 使用t-SNE聚类算法和KNN分类算法，对拉曼光谱生物医学检测数据进行处理和分析，验证实验中应用拉曼光谱仪采集乙肝感染血清及正常人血清样本的拉曼光谱数据，通过机器学习算法对拉曼光谱数据进行处理分析，验证算法对拉曼光谱实验数据处理的有效性。**结果** 利用t-SNE聚类算法和KNN分类算法进行拉曼光谱数据处理后，可以有效区分乙肝感染患者血清与对照的正常人血清。**结论** 利用拉曼光谱光谱仪采集生物医学样本光谱图像数据，通过t-SNE和KNN等机器学习算法进行处理分析，是一种可行的快速生物医学检测新方法。

[关键词] 生物医学检测；拉曼光谱；机器学习；t分布随机近邻嵌入；K最近邻算法

Research on Machine Learning Algorithms for Raman Spectroscopy Imaging for Biomedical Detection

YU Kaiming[1a], BAO Xiaodong[1b], Li Bei[2,3], HONG Xi[2], LIU Jingxin[1c]

1. a. Department of Hand Surgery; b. Medical Imaging Engineering Center; c. Department of Radiology, China-Japan Union Hospital of Jilin University, Changchun Jilin 130033, China; 2. HOOKE Instruments Ltd., Changchun Jilin 130033, China; 3. Changchun Institute of Optics, Fine Mechanics and Physics, CAS, Changchun Jilin 130033, China

Abstract: **Objective** The machine learning processing and analysis algorithms for Raman spectral images for rapid biomedical detection are discussed to provide a reference for the establishment of a convenient and fast innovative detection method for hepatitis B. **Methods** Processing and analysis of Raman spectroscopy biomedical assay data using t-SNE clustering algorithm and KNN classification algorithm. Validation experiments applied Raman spectroscopy to collect Raman spectroscopy data from hepatitis B infected serum and normal human serum samples, and machine learning algorithms were used to process and analyze the Raman spectroscopy data to verify the effectiveness of the algorithms on Raman spectroscopy experimental data processing. **Results** Raman spectral data processing using t-SNE clustering algorithm and KNN classification algorithm can effectively distinguish hepatitis B-infected sera from control normal human serum. **Conclusion** The collection of biomedical sample spectroscopy imaging data using Raman spectroscopy and processing and analysis by machine learning algorithms such as t-SNE and KNN is a viable new method for rapid biomedical detection.

Key words: biomedical testing; Raman spectroscopy; machine learning; t-SNE; KNN

[中图分类号] R318 [文献标识码] A

doi：10.3969/j.issn.1674-1633.2021.08.007 [文章编号] 1674-1633(2021)08-0026-04

引言

近年来，将拉曼光谱分析用于生物医学检测成为新的研究热点，特别是新冠疫情出现后，由于核酸检测对硬件和人员要求较高，一般医院难以完成。因此，建立便捷快速的创新检测方法对疫情防控具有重要意义，采用拉曼光谱进行快速生物医学检测成为国内外科学家探索的新方向。

拉曼光谱分析法基于拉曼散射效应，具有快速、无损、非接触的优势[1-3]，已在有机化学、高分子材料、材料科学等研究领域应用多年[4-5]。但是由于拉曼光谱数据采集处理分析时间太长，因此在医学检测领域发展缓慢。近年来，随着光学技术和计算机技术的发展，大大缩短了拉曼光谱的采集处理时间，使它应用于生物医学检测领域成为可能[6-8]。使用拉曼光谱检测时，生物样品用量很少，且无须前置处理，大大降低了操作难度，保护了样本原始性，因而可以采集到生物样品最真实的信息[9-11]。另外，拉曼光谱对于研究生物大分子的结构与性能，单细胞的核酸、蛋白质、脂质含量信息[12]以及细胞分子结构实时变化的信息等都具有显著优势[13-14]。

在使用拉曼光谱进行生物医学检测时，存在数据量大、数据维度高、光谱特征峰值多等问题。为此，需要通过计

收稿日期：2020-10-19
基金项目：国家重点研发计划（2018YFC1315604；2018YFC0116900）；吉林省科技发展计划项目（20200901017SF）；吉林大学高层次科技创新团队建设项目（2017TD-27）。
通信作者：刘景鑫，教授，主要研究方向为智能精准医学。
通信作者邮箱：jingxin@jlu.edu.cn

算机对数据进行降维和聚类分析处理，最终可以达到生物医学检测的效果[15-18]。

1 方法

由于生物检测具有复杂的环境和多样的生物，使用无监督学习在复杂乙肝血清环境中进行检测和分析，基于 t 分布随机近邻嵌入（t-Distributed Stochastic Neighbor Embedding，t-SNE）[19] 非线性拉曼光谱数据压缩，将高维拉曼光谱投影到低维平面，实现在低维空间的可视化聚类。

使用 K 最近邻算法（K-Nearest Neighbor，KNN）[20]，广泛适用于增量模型下的模式识别领域。它是一种在线学习技术，新学习样本可以直接加入训练好的训练集，而不需要重新进行学习训练从而提高了模型训练速度，且分类准确度高，对异常值的噪声有较高的容忍度，对复杂血清样品鉴别有着天生的优势。

1.1 t–SNE聚类算法

SNE 算法可以保持数据在进行降维处理前后各数据点间远近关系的概率，从而可以保持降维前后的数据内部结构。SNE 算法的基本思想：① 利用复杂度因子，选取近邻样本；② 用概率的形式将近邻样本间的欧氏距离转化成样本相似度；③ 利用相对熵目标函数算得降维后的数据表达。其中，x_i 和 x_j 间的相似度由条件概率表达，即为 x_i 选取 x_j 作为近邻的概率；其对应的嵌入子空间 y_i 和 y_j 间的相似度使用相似的表达。

应用中发现，原算法存在着低维度数据拥挤和价值方程优化困难的问题，因而在原 SNE 算法基础上又提出了基于 t 分布的 t-SNE 算法。t-SNE 算法较原算法的优点：① x_i 和 x_j 间的相似度由联合概率表达，联合概率具有对称性；② 嵌入子空间 y_i 和 y_j 间的相似度则用 t 分布表达。

设 $X=\{x_1,x_2,\cdots,x_n\}$，其中 x_i 为 d 维向量，联合概率 p_{ij} 表示数据 x_i 和 x_j 之间的相似度，即原空间中 x_i 选取 x_j 作为邻近的概率，即式 (1)：

$$p_{ij}=\frac{\exp(\frac{\|s_i-x_j\|^2}{2\lambda^2})}{\sum_{k\neq l}\exp(\frac{\|x_k-x_l\|^2}{2\lambda^2})} \tag{1}$$

其中，λ 是高斯函数的方差，$p_{ij}=0$ 数据间相似度概率总和为 1。

取 n 个 r 维向量 $Y=\{y_1,y_2,\cdots,y_n\}$（r 远小于 d），作为 X 对应的子空间数据，利用 t 分布 q_{ij} 表示子空间 y_i 和 y_j 间的相似度，即子空间数据间的概率，即式 (2)：

$$q_{ij}=\frac{(1+\|y_i-y_i\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k-y_l\|^2)^{-1}} \tag{2}$$

t-SNE 通过最小化目标函数，即式 (3)：

$$C(Y)=\sum_{i\neq j}p_{ij}lg\frac{p_{ij}}{q_{ij}} \tag{3}$$

获取最佳子空间的向量表达，即最小化原空间和子空间两个概率分布的相对熵，其本质就是最大限度地匹配 p_{ij} 和 q_{ij}，再利用梯度下降法计算式 (3) 最优值。

求解时，最优化过程中存在振荡现象，为了改善这个问题，并加快最优化过程，在式 (3) 上添加一个动量项，从而有了带动量的梯度，见式 (4)：

$$Y^{(m+1)}=Y^{(m)}+\eta\frac{dC(Y)}{d(y_i)}\Big|_{Y=Y(m)}+\beta(m)(Y^{(m)}-Y^{(m-1)}) \tag{4}$$

其中 $Y^{(m)}=[y_1^{(m)},y_2^{(m)},\cdots,y_n^{(m)}]\in R^{r\times n}$ 表示第 m 次迭代向量 Y 的值，η 为学习速率，$\beta(m)$ 为第 m 次迭代的动量值。

1.2 KNN分类算法

在 N 个训练样本中，找到测试样本 x 的 k 个近邻。设数据集中有 m 个训练样本，并有 c 个类别，即 $\{\omega_1,\cdots,\omega_c\}$，测试样本为 x。则 KNN 算法可描述为：在 m 个训练样本中找到 x 的 k 个邻域，其中 x 的 k 个近邻中属于类别 wi 的样本数户分别为 $k_1,k_2,\cdots,k_n$ 则判别函数见式 (5)：

$$g_i(x)=k_i \quad i=1,2,\cdots,c \tag{5}$$

决策规则为式 (6)：

$$g_j(x)=\max_i k_i \tag{6}$$

则决策 $x\in\omega_j$。

KNN 的思想是给出一个样本集合和一个合适的距离度量方式，对任意的一个测试样本，找到离它最近的 k 个样本，根据这 k 个样本的类别统计信息决定此测试样本的类别归属问题，即将待分样本 x 归类为与其 k 个近邻中出现次数最多的类别。KNN 算法的基本要素为：k 值，距离度量方式和分类的决策规则。

KNN 分类步骤：① 准备训练样本集 X，其中包含 n 个训练样本，根据具体要求选择一个合适的距离度量方式，用 $dis(x_a,x_b)$ 表示样本集中的 x_a、x_b 这两点的距离；② 对于测试样本 x，利用距离度量公式计算测试样本 x 与 n 个样本的距离，得到距离集合 Dis，其中 $Dis=\{dis(x,x_1),dis(x,x_1),\cdots,dis(x,x_n),\}$；③ 对距离集合进行排序，从中选择最小的 k 个元素，从而得到 k 个元素对应的 k 个样本；④ 对这 k 个样本所属类别进行统计，用投票的方式得到最终分类结果。

2 结果

为验证本研究中提出的拉曼光谱数据处理方法，我们使用乙肝感染血清及正常人血清进行了实验验证。验证实验使用了 2 位乙肝患者血清和 2 位正常人血清。每个样品取中心位置，各测试 50 组拉曼数据。样品前置处理使用离心取全血的血清（其中每组各有一份样品有轻微溶血现象，血清偏红色，有血红素干扰）；将血清用棉签点样于检测芯片上，风干后待测。在对血清进行拉曼光谱检测后，首先对数据进行预处理，进行宇宙射线、平滑、基线校正处理，再以最高峰为标准对所有光谱进行归一化，生成光谱图。

2.1 拉曼光谱测试结果

经检测，分别得到 2 位正常人血清拉曼光谱测试区域及光谱图（图 1 和图 2）和 2 位乙肝患者血清（图 3 和图 4）。

2.2 聚类算法分析结果

经过 t-SNE 算法处理后，直接可以得到聚类分析结果图（图 5）。从图 5 中可以看出，阳性组数据与对照组存在明显差异，但对照组数据的类内聚合度较低，组内差异比较大。

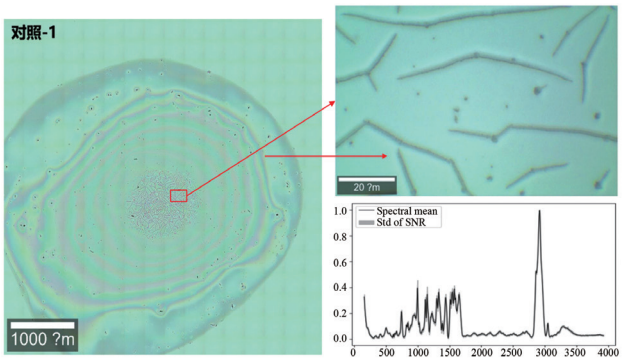

图1 对照-1实验拉曼光谱测试区域及光谱图

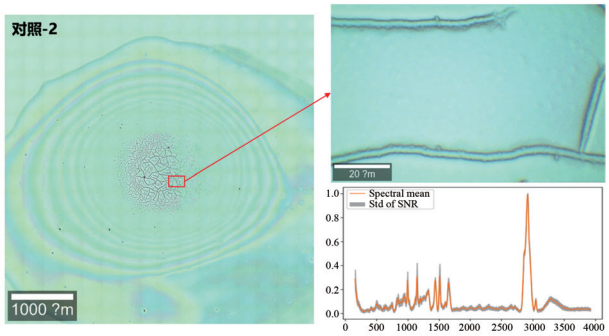

图2 对照-2实验拉曼光谱测试区域及光谱图

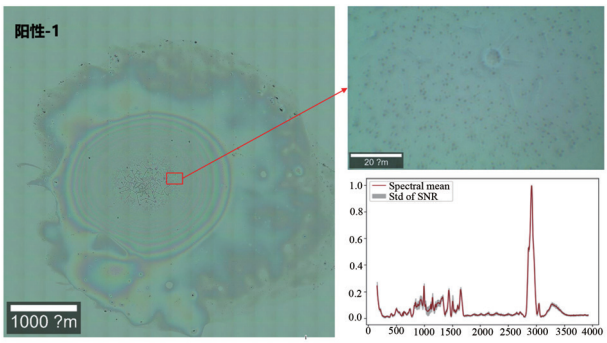

图3 阳性-1实验拉曼光谱测试区域及光谱图

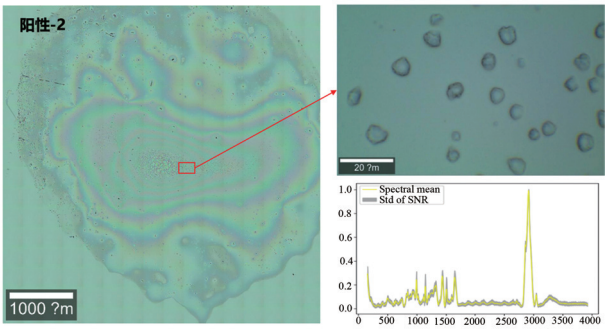

图4 阳性-2实验拉曼光谱测试区域及光谱图

2.3 分类算法分析结果

使用 KNN 分类算法，实验数据进行分类分析，得到分类分析结果图（图 6）。从图 6 中可以看出，基于目前的数据，可以根据拉曼光谱对阳性组与对照组进行区分，验证了拉曼光谱进行生物医学检测可行性及相关算法分析处理检测数据的可行性。

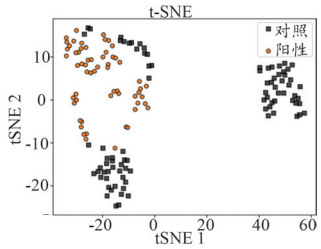

图5 聚类分析结果图

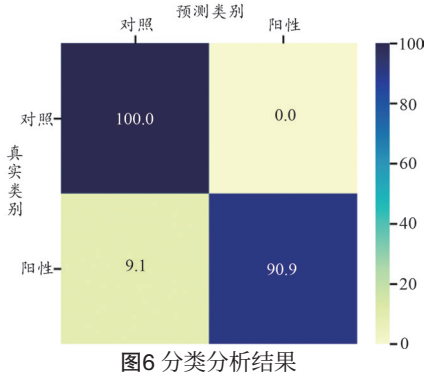

图6 分类分析结果

3 讨论

我们利用机器学习的方法，对拉曼光谱进行生物医学检测的数据进行处理分析，从乙肝感染血清验证实验的结果来看：阳性组与对照组的拉曼光谱图像存在差异，可以进行区分；从验证实验的数据分析来看：对每个样品的数据进行分析，阳性组的 2 个样品数据的类内聚合度较高，而对照组的 2 组数据差异较大。由此可见，生物的拉曼光谱图像数据是可以表征生物特性的，t-SNE 聚类算法、KNN 分类算法等机器学习算法在对生物医学拉曼光谱数据处理方面也是可行的，特别在同类组别的区分上显示出了算法的有效性。

然而，同组内个体间的差异也较为明显，数据结果受样本自身反应变化影响较多，对于此问题，后续工作将从两个方面展开：① 优化样本的采集、制作和保存，保证样本的完整性和统一性；② 项目正式开展后采集更多的样本数据，探究个体差异原因，并根据差异的特点开发相应算法，提高检测准确率。

目前，我们对于拉曼光谱用于生物医学检测的数据处理分析方法研究取得了初步的结果。同时，我们也在尝试把深度学习方法应用到拉曼光谱检测数据的处理中，以求可以更加准确高效地完成生物医疗检测，获得更高的检测结果。后续我们还需要从算法优化、样品制备、检测流程标准化等角度加以优化和完善，通过大量细菌、病毒微生物检测实验分析以提升拉曼光谱在生物医学检测领域的检测范围与检测精度。

将拉曼光谱分析应用到生物医学检测中，可以形成检测细胞、细菌甚至病毒微生物的一种新型快速便捷的检测技术，拉曼光谱也将有希望成为快速检测新型冠状病毒的新方法。

[参考文献]

[1] 李晓舟,孙宝明,杨天月,等.利用PCA-LDA和参数-CART方法对直肠癌血清拉曼光谱的分析[J].激光生物学报,2015,24(1):73-76.

[2] 张海鹏,付彤,张志茹,等.应用PCA方法分析拉曼光谱检测结果对乳腺良恶性疾病鉴别诊断的价值[J].吉林大学学报(医学版),2013,39(5):938-943.

[3] 董金颖.共聚焦显微拉曼光谱技术在病原微生物快速检测中的研究[D].大连:大连医科大学,2018.

[4] 欧阳磊.高性能表面增强拉曼基底的构建及应用[D].武汉:华中科技大学,2017.

[5] Santos MCD,Nascimento YM,Josélio M,*et al*.ATR-FTIR spectroscopy coupled with multivariate analysis techniques for the identification of DENV-3 in different concentrations in blood and serum: a new approach[J].*Rsc Advances*,2017,7(41):25640-25649

[6] Roy S,Perez-Guaita D,Bowden S,*et al*.Spectroscopy goes Viral: Diagnosis of hepatitis B and C virus infection from human sera using ATR-FTIR spectroscopy[J].*Clin Spect*,2019,(1):100001.

[7] Nawaz H,Rashid N,Saleem M,*et al*.Prediction of viral loads for diagnosis of hepatitis C infection in human plasma samples using Raman spectroscopy coupled with partial least squares regression analysis[J].*J Raman Spect*,2017,48(5):697-704.

[8] Prats MB,Harreither E,Schosserer M,*et al*.Label-free live cell imaging by Confocal Raman Microscopy identifies CHO host and producer cell lines[J].*Biotechnol J*,2017,12(1):1600037.

[9] 高健赫,蔡红星,张喜和.鼻咽癌细胞分泌物的拉曼光谱研究[J].光散射学报,2012,(2):186-192.

[10] 佟东倪.拉曼光谱技术在乙型病毒肝炎血清诊断试验中的应用[D].乌鲁木齐:新疆医科大学,2020.

[11] 廖发电.原发性肝癌患者组织和血浆球蛋白的表面增强拉曼光谱研究[D].福州:福建师范大学,2016.

[12] Schie IW,Huser T.Methods and applications of Raman microspectroscopy to single-cell analysis[J].*Appl Spect*,2013,67(8):813-828.

[13] 周涵婧.胃癌血清的表面增强拉曼光谱研究[D].青岛:中国海洋大学,2013.

[14] 王文惠,杨卫华,堵一乔,等.糖尿病患者血清拉曼光谱与常规检测的临床分析[J].实用临床医药杂志,2019,23(12):12-15.

[15] 潘琨琨,张朝霞,秦洁,等.基于血清拉曼光谱的丙型肝炎病毒诊断和1b亚型的鉴定[J].新疆医科大学学报,2019,42(5):656-662.

[16] 邵锋.三维仿生表面增强拉曼散射基底构建及其用于动物病毒非标记检测[D].武汉:华中农业大学,2014.

[17] 陈阳,严霞,张旭,等.基于支持向量机算法的多环芳烃表面增强拉曼光谱的定量分析[J].中国激光,2019,46(3):298-305.

[18] 赵恒,陈娱欣,续小丁,等.基于局部对称重加权惩罚最小二乘的拉曼基线校正[J].中国激光,2018,45(12):280-291.

[19] 彭跃辉,车辚辚.基于t-SNE的PQD特征提取可视化仿真分析[J].华北电力大学学报(自然科学版),2019,46(6):36-40.

[20] 崔浩康.多标签学习算法的改进与研究[D].成都:电子科技大学,2020.

本文编辑 崔丽君

上接第18页

of technologies and applications[J].*Diabetes Metab J*,2019,43(4):383-397.

[12] Saudek CD,Kaplan J,Boyne MS,*et al*.Timing of changes in interstitial and venous blood glucose measured with a continuous subcutaneous glucose sensor[J].*Diabetes*,2003,52(11):2790-2794.

[13] Shi T,Li D,Li G,*et al*.Modeling and measurement of correlation between blood and interstitial glucose changes[J].*J Diabetes Res*,2017,(2017):1.

[14] 李澍,郝烨,任海萍.医疗器械无线共存评价方法研究[J].中国医疗设备,2019,34(9):8-12,23.

[15] 孙于萍.基于敏感参数的高温高湿环境人体生理响应研究[D].天津:天津大学,2012.

[16] 杜晨秋.环境温度变化对人体热调节和健康影响及其分子机理研究[D].重庆:重庆大学,2018.

[17] 王倩,金娟,陈冬晓,等.温度对葡萄糖氧化酶构象及电催化活性的影响[A].2009年第十五次全国电化学学术会议论文集[C].中国化学会,2009:1-4.

[18] 陈玮.皮下植入式持续血糖监测微传感器技术研究[D].杭州:浙江大学,2017.

[19] 方璐.介入式电化学血糖监测传感器的研究[D].杭州:浙江大学,2014.

[20] 姚伟,朱彬,丁光宏,等.氧代谢的整体动力学模型及生理分析[J].中国生物医学工程学报,2008,27(1):50-55.

[21] 周竞,颜巍,施建丰,等.维生素C对3种血糖检测方法的干扰效应分析[J].国际检验医学杂志,2017,38(10):1336-1337.

本文编辑 崔丽君